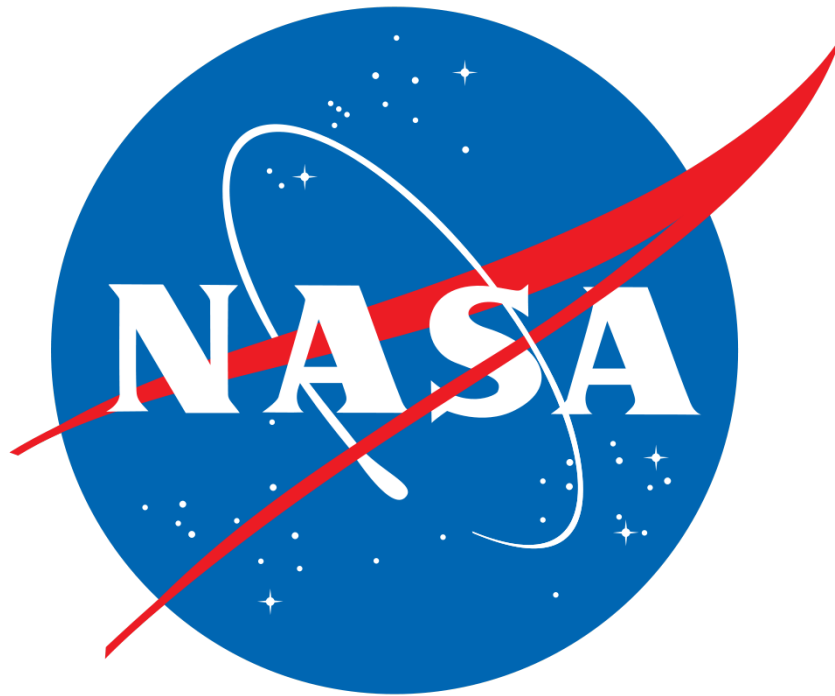


Modelización del Radio Planetario en Exoplanetas: Un Enfoque de Regresión Lineal Múltiple

Fernando José Pereira Salvador

2026-05-10



Contents

1	Introducción	3
1.1	Descripción de la base de datos	3
1.2	Selección y justificación de variables	9
1.2.1	Variables no incorporadas al modelo	9
1.2.2	Variables seleccionadas para el análisis	10
2	Análisis Exploratorio de Datos	12
2.1	Resumen descriptivo	12
2.2	Análisis gráfico	13
2.3	Correlaciones con la variable dependiente	14
2.4	Análisis previo de multicolinealidad	16
2.5	Correlaciones parciales con la variable dependiente	18
2.6	Conclusión del análisis exploratorio	19
3	Ajuste del modelo lineal	20
3.1	Conclusión	24
3.2	Recomendaciones	26
4	Bondad del ajuste	27
5	Análisis de los residuos	28
5.1	Valores sobreinfluyentes	33
6	Selección de variables	35
7	Conclusiones finales	43
8	Ampliación: modelo con transformación logarítmica	44
8.1	Comparación de la bondad del ajuste	45
8.2	Intervalos de confianza del modelo logarítmico	46
8.3	Diagnóstico de residuos del modelo logarítmico	46
8.4	Conclusión de la ampliación	48

1 Introducción

1.1 Descripción de la base de datos

La base de datos utilizada procede del **NASA Exoplanet Archive**, concretamente de la tabla *Planetary Systems Composite Parameters*. Esta base recoge información sobre exoplanetas confirmados y sus sistemas planetarios, incluyendo variables relativas al planeta, a su órbita, a la estrella anfitriona y a la distancia del sistema respecto a la Tierra.

El objetivo del análisis es estudiar qué características físicas, orbitales y estelares están asociadas con el tamaño de un exoplaneta. Para ello, se selecciona como variable dependiente el radio planetario (`pl_rade`), medido en radios terrestres. Esta variable permite analizar el tamaño relativo de los exoplanetas respecto a la Tierra.

Desde el punto de vista del problema, el modelo puede entenderse como una herramienta exploratoria para estudiar qué factores están relacionados con el tamaño planetario. Este tipo de análisis puede ser útil en el contexto de la investigación astronómica, por ejemplo, para caracterizar poblaciones de exoplanetas, identificar patrones físicos y apoyar la priorización de candidatos para observaciones posteriores.

```
library(ppcor)
```

```
## Cargando paquete requerido: MASS
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.1      v stringr    1.5.1
## v ggplot2    4.0.2      v tibble     3.3.0
## v lubridate  1.9.4      v tidyr      1.3.1
## v purrr      1.1.0
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(broom)
library(tibble)
library(lmtest)
```

```
## Cargando paquete requerido: zoo
##
## Adjuntando el paquete: 'zoo'
##
## The following objects are masked from 'package:base':
##
##      as.Date, as.Date.numeric
```

```
exoplanetas <- read_csv(
  "PSCompPars_2026.04.29_08.19.44.csv",
  comment = "#"
)
```

```
## Rows: 6273 Columns: 84
## -- Column specification -----
## Delimiter: ","
## chr (9): pl_name, hostname, discoverymethod, disc_facility, pl_bmassprov, s...
## dbl (75): sy_snum, sy_pnum, disc_year, pl_controv_flag, pl_orbper, pl_orbper...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
glimpse(exoplanetas)
```

```
## Rows: 6,273
## Columns: 84
## $ pl_name      <chr> "11 Com b", "11 UMi b", "14 And b", "14 Her b", "16 Cy~
## $ hostname     <chr> "11 Com", "11 UMi", "14 And", "14 Her", "16 Cyg B", "1~
## $ sy_snum      <dbl> 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, ~
## $ sy_pnum      <dbl> 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, ~
## $ discoverymethod <chr> "Radial Velocity", "Radial Velocity", "Radial Velocity~
## $ disc_year     <dbl> 2007, 2009, 2008, 2002, 1996, 2020, 2008, 2018, ~
## $ disc_facility <chr> "Xinglong Station", "Thueringer Landessternwarte Taute~
## $ pl_controv_flag <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pl_orbper     <dbl> 323.21, 516.22, 186.76, 1766.41, 798.50, 578.38, 982.8~
## $ pl_orbpererr1 <dbl> 0.06, 3.20, 0.11, 0.67, 1.00, 2.01, 1.06, NA, 0.00, 2.~
## $ pl_orbpererr2 <dbl> -0.05, -3.20, -0.12, -0.68, -1.00, -2.09, -0.92, NA, --
## $ pl_orbperlim  <dbl> 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, NA, NA, NA, ~
## $ pl_orbsmax    <dbl> 1.178, 1.530, 0.775, 2.839, 1.660, 1.450, 2.476, 330.0~
## $ pl_orbsmaxerr1 <dbl> 0.000, 0.070, 0.000, 0.039, 0.030, 0.020, 0.002, NA, 0~
## $ pl_orbsmaxerr2 <dbl> 0.000, -0.070, 0.000, -0.041, -0.030, -0.020, -0.002, ~
## $ pl_orbsmaxlim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pl_rade      <dbl> 12.20000, 12.30000, 13.10000, 12.50000, 13.50000, 12.9~
## $ pl_radeerr1   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_radeerr2   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_radelim    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pl_radadj     <dbl> 1.090, 1.090, 1.160, 1.120, 1.200, 1.150, 1.120, 1.664~
## $ pl_radjerr1   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_radjerr2   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_radjlim    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pl_bmasse     <dbl> 4914.8985, 4684.8142, 1131.1513, 2828.6728, 565.7374, ~
## $ pl_bmasseerr1 <dbl> 39.09289, 794.57500, 36.23244, 413.17693, 25.42640, 47~
## $ pl_bmasseerr2 <dbl> -39.72855, -794.57500, -38.77507, -540.30829, -25.4264~
## $ pl_bmasselim  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pl_bmassj     <dbl> 15.464, 14.740, 3.559, 8.900, 1.780, 4.320, 9.207, 8.0~
## $ pl_bmassjerr1 <dbl> 0.123, 2.500, 0.114, 1.300, 0.080, 0.150, 0.160, 1.000~
## $ pl_bmassjerr2 <dbl> -0.125, -2.500, -0.122, -1.700, -0.080, -0.120, -0.077~
## $ pl_bmassjlim  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pl_bmassprov  <chr> "Msini", "Msini", "Msini", "Mass", "Msini", "Msini", "~
## $ pl_orbeccen   <dbl> 0.2380, 0.0800, 0.0000, 0.3683, 0.6800, 0.0600, 0.0240~
## $ pl_orbeccenerr1 <dbl> 0.0070, 0.0300, NA, 0.0029, 0.0200, 0.0300, 0.0070, NA~
```

```

## $ pl_orbeccenerr2 <dbl> -0.0070, -0.0300, NA, -0.0029, -0.0200, -0.0200, -0.01~
## $ pl_orbeccenlim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, NA, NA, NA, NA, ~
## $ pl_insol <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_insolerr1 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_insolerr2 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_insollim <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ pl_eqt <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 1700, NA, NA, NA, 1450, NA~
## $ pl_eqterr1 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 100, NA, NA, NA, 50, NA, 1~
## $ pl_eqterr2 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, -100, NA, NA, NA, -50, NA, ~
## $ pl_eqtlim <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, NA, 0, NA, 0, N~
## $ ttv_flag <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ st_spectype <chr> "G8 III", "K4 III", "K0 III", "K0", "G3 V", "K3 III", ~
## $ st_teff <dbl> 4874, 4213, 4888, 5338, 5750, 4157, 4980, 4060, 4816, ~
## $ st_tefferr1 <dbl> NA, 46, NA, 25, 8, 11, NA, 300, NA, 44, 44, 100, NA, 5~
## $ st_tefferr2 <dbl> NA, -46, NA, -25, -8, -10, NA, -200, NA, -44, -44, -10~
## $ st_tefflim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, ~
## $ st_rad <dbl> 13.760000, 29.790000, 11.550000, 0.930000, 1.130000, 2~
## $ st_raderr1 <dbl> 2.850000, 2.840000, 1.120000, 0.010000, 0.010000, 0.78~
## $ st_raderr2 <dbl> -2.450000, -2.840000, -0.510000, -0.010000, -0.010000, ~
## $ st_radlim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, ~
## $ st_mass <dbl> 2.090000, 2.780000, 1.780000, 0.970000, 1.080000, 1.22~
## $ st_masserr1 <dbl> 0.640000, 0.690000, 0.430000, 0.040000, 0.040000, 0.13~
## $ st_masserr2 <dbl> -0.630000, -0.690000, -0.290000, -0.040000, -0.040000, ~
## $ st_masslim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ st_met <dbl> -0.260, -0.020, -0.210, 0.430, 0.060, -0.010, -0.060, ~
## $ st_meterr1 <dbl> 0.10, NA, 0.10, 0.07, NA, 0.10, 0.10, NA, 0.10, 0.04, ~
## $ st_meterr2 <dbl> -0.10, NA, -0.10, -0.07, NA, -0.10, -0.10, NA, -0.10, ~
## $ st_metlim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, NA, NA, NA, NA, N~
## $ st_metratio <chr> "[Fe/H]", "[Fe/H]", "[Fe/H]", "[Fe/H]", "[Fe/H]", "[Fe~
## $ st_logg <dbl> 2.45000, 1.93000, 2.55000, 4.45000, 4.36000, 1.70000, ~
## $ st_loggerr1 <dbl> 0.080000, 0.070000, 0.060000, 0.020000, 0.010000, 0.04~
## $ st_loggerr2 <dbl> -0.080000, -0.070000, -0.070000, -0.020000, -0.010000, ~
## $ st_logglim <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, ~
## $ rastr <chr> "12h20m42.91s", "15h17m05.90s", "23h31m17.80s", "16h10~
## $ ra <dbl> 185.17878, 229.27460, 352.82415, 242.60210, 295.46564, ~
## $ decstr <chr> "+17d47m35.71s", "+71d49m26.19s", "+39d14m09.01s", "+4~
## $ dec <dbl> 17.7932516, 71.8239428, 39.2358367, 43.8163621, 50.516~
## $ sy_dist <dbl> 93.18460, 125.32100, 75.43920, 17.93230, 21.13970, 124~
## $ sy_disterr1 <dbl> 1.9238000, 1.9765000, 0.7140000, 0.0073000, 0.0110000, ~
## $ sy_disterr2 <dbl> -1.9238000, -1.9765000, -0.7140000, -0.0073000, -0.011~
## $ sy_vmag <dbl> 4.72307, 5.01300, 5.23133, 6.61935, 6.21500, 5.22606, ~
## $ sy_vmagerr1 <dbl> 0.023, 0.005, 0.023, 0.023, 0.016, 0.023, 0.023, 0.069~
## $ sy_vmagerr2 <dbl> -0.023, -0.005, -0.023, -0.023, -0.016, -0.023, -0.023~
## $ sy_kmag <dbl> 2.282000, 1.939000, 2.331000, 4.714000, 4.651000, 2.09~
## $ sy_kmagerr1 <dbl> 0.346, 0.270, 0.240, 0.016, 0.016, 0.244, 0.204, 0.021~
## $ sy_kmagerr2 <dbl> -0.346, -0.270, -0.240, -0.016, -0.016, -0.244, -0.204~
## $ sy_gaiamag <dbl> 4.44038, 4.56216, 4.91781, 6.38300, 6.06428, 4.75429, ~
## $ sy_gaiamagerr1 <dbl> 0.0038479, 0.0039035, 0.0028262, 0.0003512, 0.0006029, ~
## $ sy_gaiamagerr2 <dbl> -0.0038479, -0.0039035, -0.0028262, -0.0003512, -0.000~

```

[summary\(exoplanetas\)](#)

##	pl_name	hostname	sy_snum	sy_pnum
##	Length:6273	Length:6273	Min. :1.000	Min. :1.000

```

## Class :character      Class :character      1st Qu.:1.000      1st Qu.:1.000
## Mode :character      Mode :character      Median :1.000      Median :1.000
##                               Mean :1.102      Mean :1.759
##                               3rd Qu.:1.000      3rd Qu.:2.000
##                               Max. :4.000      Max. :8.000
##
## discoverymethod      disc_year      disc_facility      pl_controv_flag
## Length:6273      Min. :1992      Length:6273      Min. :0.000000
## Class :character      1st Qu.:2014      Class :character      1st Qu.:0.000000
## Mode :character      Median :2016      Mode :character      Median :0.000000
##                               Mean :2017      Mean :0.007492
##                               3rd Qu.:2021      3rd Qu.:0.000000
##                               Max. :2026      Max. :1.000000
##                               NA's :1
## pl_orbper      pl_orbpererr1      pl_orbpererr2
## Min. : 0      Min. : 0      Min. : -100000000
## 1st Qu.: 4      1st Qu.: 0      1st Qu.: 0
## Median : 11      Median : 0      Median : 0
## Mean : 72131      Mean : 87931      Mean : -20085
## 3rd Qu.: 38      3rd Qu.: 0      3rd Qu.: 0
## Max. :402000000      Max. :470000000      Max. : 0
## NA's :339      NA's :829      NA's :829
## pl_orbperlim      pl_orbsmax      pl_orbsmaxerr1
## Min. : -1.00000      Min. :4.400e-03      Min. : 0.0000
## 1st Qu.: 0.00000      1st Qu.:5.230e-02      1st Qu.: 0.0007
## Median : 0.00000      Median :1.020e-01      Median : 0.0020
## Mean : -0.00101      Mean :1.569e+01      Mean : 1.8045
## 3rd Qu.: 0.00000      3rd Qu.:3.070e-01      3rd Qu.: 0.0160
## Max. : 0.00000      Max. :1.900e+04      Max. :5205.0000
## NA's :339      NA's :428      NA's :2367
## pl_orbsmaxerr2      pl_orbsmaxlim      pl_rade      pl_radeerr1
## Min. : -2.060e+03      Min. : -1.00000      Min. : 0.3098      Min. : 0.0000
## 1st Qu.: -1.600e-02      1st Qu.: 0.00000      1st Qu.: 1.8400      1st Qu.: 0.1300
## Median : -2.000e-03      Median : 0.00000      Median : 2.8414      Median : 0.2700
## Mean : -9.608e-01      Mean : -0.00051      Mean : 5.8064      Mean : 0.5147
## 3rd Qu.: -7.300e-04      3rd Qu.: 0.00000      3rd Qu.:11.9000      3rd Qu.: 0.5500
## Max. : 0.000e+00      Max. : 0.00000      Max. :87.2059      Max. :68.9100
## NA's :2367      NA's :427      NA's :50      NA's :1933
## pl_radeerr2      pl_radelim      pl_radj      pl_radjerr1
## Min. : -32.5061      Min. : -1.00000      Min. :0.02764      Min. :0.00000
## 1st Qu.: -0.4484      1st Qu.: 0.00000      1st Qu.:0.16415      1st Qu.:0.01158
## Median : -0.2200      Median : 0.00000      Median :0.25350      Median :0.02409
## Mean : -0.4136      Mean : -0.000643      Mean :0.51804      Mean :0.04592
## 3rd Qu.: -0.1100      3rd Qu.: 0.00000      3rd Qu.:1.06000      3rd Qu.:0.04907
## Max. : 0.0000      Max. : 0.00000      Max. :7.78000      Max. :6.14774
## NA's :1933      NA's :50      NA's :50      NA's :1933
## pl_radjerr2      pl_radjlim      pl_bmasse      pl_bmasseerr1
## Min. : -2.90000      Min. : -1.00000      Min. : 0.02      Min. : 0.00
## 1st Qu.: -0.04000      1st Qu.: 0.00000      1st Qu.: 4.27      1st Qu.: 2.20
## Median : -0.01963      Median : 0.00000      Median : 9.20      Median : 18.43
## Mean : -0.03691      Mean : -0.000643      Mean : 400.73      Mean : 177.06
## 3rd Qu.: -0.00981      3rd Qu.: 0.00000      3rd Qu.: 184.02      3rd Qu.: 79.46
## Max. : 0.00000      Max. : 0.00000      Max. :9534.85      Max. :12652.75
## NA's :1933      NA's :50      NA's :31      NA's :3272

```

```

## pl_bmasseerr2      pl_bmasselim      pl_bmassj      pl_bmassjerr1
## Min.      :-6038.74  Min.      :-1.0000  Min.      :6.290e-05  Min.      : 0.0000
## 1st Qu.:  -69.92  1st Qu.: 0.0000  1st Qu.:1.340e-02  1st Qu.: 0.0070
## Median :  -17.00  Median : 0.0000  Median :2.895e-02  Median : 0.0580
## Mean   : -127.57  Mean   : 0.0322  Mean   :1.261e+00  Mean   : 0.5575
## 3rd Qu.:  -2.07  3rd Qu.: 0.0000  3rd Qu.:5.790e-01  3rd Qu.: 0.2500
## Max.    :   0.00  Max.    : 1.0000  Max.    :3.000e+01  Max.    :39.8100
## NA's    :3272    NA's    :31      NA's    :31      NA's    :3272
## pl_bmassjerr2      pl_bmassjlim      pl_bmassprov      pl_orbeccen
## Min.      :-19.0000  Min.      :-1.0000  Length:6273  Min.      :0.0000
## 1st Qu.:  -0.2200  1st Qu.: 0.0000  Class :character  1st Qu.:0.0000
## Median :  -0.0530  Median : 0.0000  Mode  :character  Median :0.0000
## Mean   :  -0.4015  Mean   : 0.0322                      Mean   :0.0791
## 3rd Qu.:  -0.0065  3rd Qu.: 0.0000                      3rd Qu.:0.0900
## Max.    :   0.0000  Max.    : 1.0000                      Max.    :0.9500
## NA's    :3272    NA's    :31      NA's    :1052
## pl_orbeccenerr1    pl_orbeccenerr2    pl_orbeccenlim      pl_insol
## Min.      :0.0000  Min.      :-0.7030  Min.      :-1.00000  Min.      : 0.00
## 1st Qu.:0.0200  1st Qu.: -0.0780  1st Qu.: 0.00000  1st Qu.: 24.11
## Median :0.0480  Median : -0.0400  Median : 0.00000  Median : 100.00
## Mean   :0.0676  Mean   : -0.0548  Mean   : 0.05229  Mean   : 419.89
## 3rd Qu.:0.0900  3rd Qu.: -0.0180  3rd Qu.: 0.00000  3rd Qu.: 376.29
## Max.    :0.5000  Max.    : 0.0000  Max.    : 1.00000  Max.    :44900.00
## NA's    :4403    NA's    :4403    NA's    :1052    NA's    :1876
## pl_insolerr1      pl_insolerr2      pl_insolllim      pl_eqt
## Min.      : 0.000  Min.      :-7200.000  Min.      :0      Min.      : 34.0
## 1st Qu.: 1.345  1st Qu.: -22.020  1st Qu.:0      1st Qu.: 569.0
## Median : 5.905  Median : -5.215  Median :0      Median : 823.0
## Mean   : 41.362  Mean   : -32.048  Mean   :0      Mean   : 914.8
## 3rd Qu.: 26.465  3rd Qu.: -1.252  3rd Qu.:0      3rd Qu.:1164.0
## Max.    :8100.000  Max.    : 0.000  Max.    :0      Max.    :4050.0
## NA's    :2633    NA's    :2633    NA's    :1876    NA's    :1600
## pl_eqtterr1      pl_eqtterr2      pl_eqtlim      ttv_flag
## Min.      : 0.00  Min.      :-1217.00  Min.      :0.00000  Min.      :0.00000
## 1st Qu.: 11.25  1st Qu.: -34.00  1st Qu.:0.00000  1st Qu.:0.00000
## Median : 20.00  Median : -19.21  Median :0.00000  Median :0.00000
## Mean   : 31.50  Mean   : -31.00  Mean   :0.00064  Mean   :0.07795
## 3rd Qu.: 35.00  3rd Qu.: -11.00  3rd Qu.:0.00000  3rd Qu.:0.00000
## Max.    :1217.00  Max.    : -0.32  Max.    :1.00000  Max.    :1.00000
## NA's    :4458    NA's    :4458    NA's    :1600
## st_spectype      st_teff      st_tefferr1      st_tefferr2
## Length:6273      Min.      : 415  Min.      : 1.00  Min.      : -2360.24
## Class :character  1st Qu.: 4896  1st Qu.: 59.76  1st Qu.: -124.62
## Mode  :character  Median : 5543  Median : 86.00  Median : -86.67
##                  Mean   : 5393  Mean   : 98.49  Mean   : -99.26
##                  3rd Qu.: 5897  3rd Qu.: 120.33  3rd Qu.: -58.33
##                  Max.    :57000  Max.    :1763.10  Max.    : -1.00
##                  NA's    :294    NA's    :608    NA's    :608
## st_tefflim      st_rad      st_raderr1      st_raderr2
## Min.      :0.000000  Min.      : 0.0115  Min.      : 0.0000  Min.      : -28.9700
## 1st Qu.:0.000000  1st Qu.: 0.7700  1st Qu.: 0.0230  1st Qu.: -0.1050
## Median :0.000000  Median : 0.9500  Median : 0.0500  Median : -0.0500
## Mean   :0.000167  Mean   : 1.4921  Mean   : 0.1880  Mean   : -0.1469
## 3rd Qu.:0.000000  3rd Qu.: 1.2400  3rd Qu.: 0.1648  3rd Qu.: -0.0230

```

```

## Max. :1.000000 Max. :88.4750 Max. :104.5280 Max. : 0.0000
## NA's :294 NA's :318 NA's :591 NA's :592
## st_radlim st_mass st_masserr1 st_masserr2
## Min. :0 Min. : 0.0094 Min. : 0.00020 Min. : -10.02000
## 1st Qu.:0 1st Qu.: 0.7700 1st Qu.: 0.02900 1st Qu.: -0.08000
## Median :0 Median : 0.9400 Median : 0.04800 Median : -0.04900
## Mean :0 Mean : 0.9355 Mean : 0.08531 Mean : -0.07901
## 3rd Qu.:0 3rd Qu.: 1.0822 3rd Qu.: 0.08900 3rd Qu.: -0.02900
## Max. :0 Max. :10.9400 Max. :10.02000 Max. : 0.00000
## NA's :318 NA's :9 NA's :283 NA's :284
## st_masslim st_met st_meterr1 st_meterr2
## Min. :0 Min. : -1.00000 Min. :0.00100 Min. : -1.0000
## 1st Qu.:0 1st Qu.: -0.08000 1st Qu.:0.04000 1st Qu.: -0.1500
## Median :0 Median : 0.02000 Median :0.08200 Median : -0.0820
## Mean :0 Mean : 0.01654 Mean :0.09657 Mean : -0.1023
## 3rd Qu.:0 3rd Qu.: 0.13000 3rd Qu.:0.14000 3rd Qu.: -0.0400
## Max. :0 Max. : 0.79000 Max. :0.96000 Max. : 0.0000
## NA's :9 NA's :640 NA's :923 NA's :923
## st_metlim st_metratio st_logg st_loggerr1
## Min. : -1 Length:6273 Min. :0.541 Min. :0.00000
## 1st Qu.: 0 Class :character 1st Qu.:4.294 1st Qu.:0.02900
## Median : 0 Mode :character Median :4.458 Median :0.05600
## Mean : 0 Mean :4.386 Mean :0.08259
## 3rd Qu.: 0 3rd Qu.:4.580 3rd Qu.:0.10300
## Max. : 1 Max. :8.070 Max. :1.10000
## NA's :640 NA's :322 NA's :614
## st_loggerr2 st_logglim rastr ra
## Min. : -3.51000 Min. : -1.00000 Length:6273 Min. : 0.1856
## 1st Qu.: -0.15000 1st Qu.: 0.00000 Class :character 1st Qu.:169.4475
## Median : -0.07300 Median : 0.00000 Mode :character Median :284.4722
## Mean : -0.09704 Mean : -0.00101 Mean :231.3077
## 3rd Qu.: -0.03000 3rd Qu.: 0.00000 3rd Qu.:293.1312
## Max. : 0.00000 Max. : 0.00000 Max. :359.9750
## NA's :614 NA's :322
## decstr dec sy_dist sy_disterr1
## Length:6273 Min. : -89.47 Min. : 1.301 Min. :3.400e-04
## Class :character 1st Qu.: -12.02 1st Qu.: 101.694 1st Qu.:4.290e-01
## Mode :character Median : 39.04 Median : 364.030 Median :3.809e+00
## Mean : 17.75 Mean : 705.406 Mean :5.927e+01
## 3rd Qu.: 45.41 3rd Qu.: 826.265 3rd Qu.:1.670e+01
## Max. : 88.83 Max. :8500.000 Max. :2.600e+03
## NA's :27 NA's :135
## sy_disterr2 sy_vmag sy_vmagerr1 sy_vmagerr2
## Min. : -2.840e+03 Min. : 0.872 Min. :0.00100 Min. : -11.92000
## 1st Qu.: -1.623e+01 1st Qu.:10.697 1st Qu.:0.03000 1st Qu.: -0.12600
## Median : -3.747e+00 Median :13.194 Median :0.06900 Median : -0.06900
## Mean : -6.649e+01 Mean :12.543 Mean :0.09732 Mean : -0.09878
## 3rd Qu.: -4.270e-01 3rd Qu.:14.918 3rd Qu.:0.12600 3rd Qu.: -0.03000
## Max. : -3.500e-04 Max. :44.610 Max. :3.10000 Max. : -0.00100
## NA's :135 NA's :299 NA's :307 NA's :312
## sy_kmag sy_kmagerr1 sy_kmagerr2 sy_gaiamag
## Min. : -3.044 Min. :0.01100 Min. : -9.99500 Min. : 2.364
## 1st Qu.: 8.406 1st Qu.:0.02000 1st Qu.: -0.03000 1st Qu.:10.420
## Median :11.060 Median :0.02300 Median : -0.02300 Median :12.925

```

## Mean	:10.374	Mean	:0.04101	Mean	:-0.04102	Mean	:12.251
## 3rd Qu.	:12.708	3rd Qu.	:0.03000	3rd Qu.	:-0.02000	3rd Qu.	:14.661
## Max.	:33.110	Max.	:9.99500	Max.	:-0.01100	Max.	:20.186
## NA's	:286	NA's	:323	NA's	:334	NA's	:343
## sy_gaiamagerr1		sy_gaiamagerr2					
## Min.	:0.00011	Min.	:-0.06323				
## 1st Qu.	:0.00026	1st Qu.	:-0.00054				
## Median	:0.00036	Median	:-0.00036				
## Mean	:0.00064	Mean	:-0.00064				
## 3rd Qu.	:0.00054	3rd Qu.	:-0.00026				
## Max.	:0.06323	Max.	:-0.00011				
## NA's	:343	NA's	:343				

1.2 Selección y justificación de variables

La base de datos original contiene 84 variables. Sin embargo, no todas se incorporan al modelo de regresión, ya que algunas tienen un papel meramente identificativo, otras representan errores de medición, otras son variables equivalentes expresadas en distintas unidades y otras tienen un elevado número de valores perdidos o una interpretación menos directa para el objetivo del análisis.

El objetivo del estudio es explicar el radio planetario (`pl_rade`), por lo que se seleccionan variables con una relación física, orbital, estelar u observacional razonable con el tamaño del planeta.

1.2.1 Variables no incorporadas al modelo

En primer lugar, se excluyen las variables identificativas o descriptivas, como `pl_name`, `hostname`, `disc_facility`, `rastr` y `decstr`. Estas variables permiten identificar el planeta, la estrella anfitriona, la instalación de descubrimiento o la posición astronómica en formato textual, pero no aportan una magnitud cuantitativa directamente interpretable en un modelo de regresión lineal.

También se excluyen variables que representan la misma magnitud en otra unidad. Por ejemplo, `pl_radj` mide el radio planetario en radios de Júpiter, mientras que `pl_rade` lo mide en radios terrestres. Como `pl_rade` se ha elegido como variable dependiente, incluir `pl_radj` como explicativa supondría introducir una variable prácticamente equivalente a la respuesta. Del mismo modo, `pl_bmassj` mide la masa planetaria en masas de Júpiter, mientras que `pl_bmasse` la mide en masas terrestres; por ello se conserva únicamente `pl_bmasse`.

Además, se excluyen las variables asociadas a errores de medición, como `pl_orbpererr1`, `pl_orbpererr2`, `pl_orbsmaxerr1`, `pl_orbsmaxerr2`, `pl_radeerr1`, `pl_radeerr2`, `pl_bmasseerr1`, `pl_bmasseerr2`, `st_tefferr1`, `st_tefferr2`, `st_raderr1`, `st_raderr2`, `st_masserr1`, `st_masserr2`, `sy_disterr1` y `sy_disterr2`. Estas columnas informan sobre la incertidumbre de las estimaciones, pero no representan directamente características físicas del planeta o de la estrella. Por este motivo, no se incluyen en el modelo principal.

De forma similar, se excluyen las variables terminadas en `lim`, como `pl_orbperlim`, `pl_orbsmaxlim`, `pl_radelim`, `pl_bmasselim`, `pl_orbeccenlim`, `pl_eqtlim`, `st_tefflim`, `st_radlim`, `st_masslim`, `st_metlim` y `st_logglim`. Estas variables indican si una medición corresponde a un límite superior o inferior, por lo que tienen un carácter auxiliar y no se interpretan como variables explicativas principales.

Por último, algunas variables potencialmente interesantes, como `pl_orbeccen`, `pl_insol`, `st_met`, `st_logg`, `sy_vmag`, `sy_kmag` o `sy_gaiamag`, no se incorporan al modelo inicial para evitar una pérdida excesiva de observaciones por valores perdidos y para mantener un modelo interpretable. No obstante, podrían considerarse en análisis complementarios o ampliaciones posteriores.

1.2.2 Variables seleccionadas para el análisis

Tras la revisión anterior, se selecciona un conjunto reducido de variables con interpretación clara y relación potencial con el radio planetario. La variable dependiente será `pl_rade`, que representa el radio del planeta en radios terrestres.

Las variables explicativas seleccionadas inicialmente son las siguientes:

Variable	Tipo	Unidad	Significado	Justificación
<code>disc_year</code>	Cuantitativa discreta	Año	Año en el que se descubrió el exoplaneta.	Puede reflejar cambios temporales en los métodos de detección y en la capacidad tecnológica para descubrir y caracterizar exoplanetas.
<code>pl_bmasse</code>	Cuantitativa continua	Masas terrestres	Masa estimada del planeta expresada en relación con la masa de la Tierra.	La masa planetaria puede estar relacionada con el tamaño del planeta, aunque ambas magnitudes no son equivalentes.
<code>pl_orbper</code>	Cuantitativa continua	Días	Tiempo que tarda el planeta en completar una órbita alrededor de su estrella.	El periodo orbital puede reflejar la posición dinámica del planeta dentro del sistema.
<code>pl_orbsmax</code>	Cuantitativa continua	Unidades astronómicas	Distancia media aproximada entre el planeta y su estrella a lo largo de su órbita.	El semieje mayor mide la distancia media aproximada entre el planeta y su estrella.
<code>pl_eqt</code>	Cuantitativa continua	Kelvin	Temperatura de equilibrio estimada del planeta.	La temperatura de equilibrio puede estar relacionada con las condiciones físicas y atmosféricas del planeta.
<code>st_teff</code>	Cuantitativa continua	Kelvin	Temperatura efectiva de la estrella anfitriona.	La temperatura de la estrella anfitriona puede influir en las condiciones térmicas del planeta.

Variable	Tipo	Unidad	Significado	Justificación
st_rad	Cuantitativa continua	Radios solares	Radio de la estrella anfitriona expresado en relación con el radio del Sol.	El radio estelar describe el tamaño de la estrella anfitriona y puede estar asociado a las características del sistema.
st_mass	Cuantitativa continua	Masas solares	Masa de la estrella anfitriona expresada en relación con la masa del Sol.	La masa estelar está relacionada con la dinámica orbital y las condiciones generales del sistema planetario.
sy_dist	Cuantitativa continua	Pársecs	Distancia entre el sistema planetario observado y la Tierra.	La distancia al sistema puede reflejar posibles sesgos observacionales en la detección y caracterización de exoplanetas.

```

datos <- exoplanetas %>%
  dplyr::select(
    disc_year,
    pl_rade,
    pl_bmasse,
    pl_orbper,
    pl_orbsmax,
    pl_eqt,
    st_teff,
    st_rad,
    st_mass,
    sy_dist
  )

datos <- datos %>%
  drop_na() %>%
  filter(
    pl_rade > 0,
    pl_bmasse > 0,
    pl_orbper > 0,
    pl_orbsmax > 0,
    sy_dist > 0
  )

```

2 Análisis Exploratorio de Datos

2.1 Resumen descriptivo

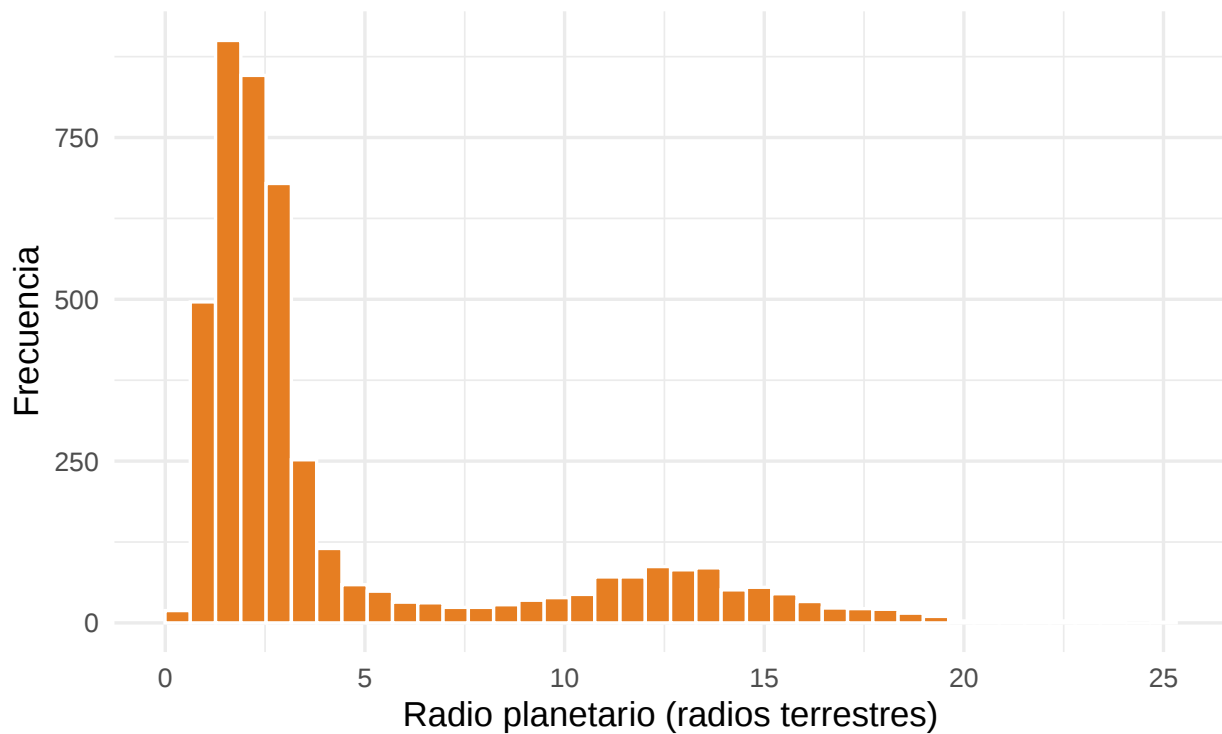
```
summary(datos)
```

```
##      disc_year      pl_rade      pl_bmasse      pl_orbper
## Min.      :1996    Min.      : 0.3098    Min.      : 0.0374    Min.      :      0
## 1st Qu.:2014    1st Qu.: 1.6400    1st Qu.: 3.5300    1st Qu.:      4
## Median :2016    Median : 2.4700    Median : 7.1422    Median :      8
## Mean      :2017    Mean      : 4.4750    Mean      :145.3384    Mean      : 93906
## 3rd Qu.:2021    3rd Qu.: 4.1975    3rd Qu.: 23.6900    3rd Qu.:     21
## Max.      :2026    Max.      :25.0000    Max.      :8899.1954    Max.      :402000000
##      pl_orbsmax      pl_eqt      st_teff      st_rad
## Min.      : 0.0050    Min.      : 55.9    Min.      : 2375    Min.      :0.0131
## 1st Qu.: 0.0474    1st Qu.: 564.0    1st Qu.: 5018    1st Qu.:0.7700
## Median : 0.0780    Median : 812.0    Median : 5598    Median :0.9400
## Mean      : 1.9882    Mean      : 907.3    Mean      : 5382    Mean      :1.0083
## 3rd Qu.: 0.1428    3rd Qu.:1159.8    3rd Qu.: 5931    3rd Qu.:1.1900
## Max.      :7506.0000    Max.      :4050.0    Max.      :10170    Max.      :6.3000
##      st_mass      sy_dist
## Min.      :0.0160    Min.      : 1.301
## 1st Qu.:0.8000    1st Qu.: 202.585
## Median :0.9403    Median : 492.485
## Mean      :0.9343    Mean      : 589.504
## 3rd Qu.:1.0800    3rd Qu.: 864.331
## Max.      :2.7800    Max.      :3460.510
```

```
ggplot(datos, aes(x = pl_rade)) +
  geom_histogram(bins = 40, fill = "#E67E22", color = "white") +
  labs(
    title = "Distribución del radio planetario",
    subtitle = "Variable dependiente del modelo",
    x = "Radio planetario (radios terrestres)",
    y = "Frecuencia"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5)
  )
```

Distribución del radio planetario

Variable dependiente del modelo



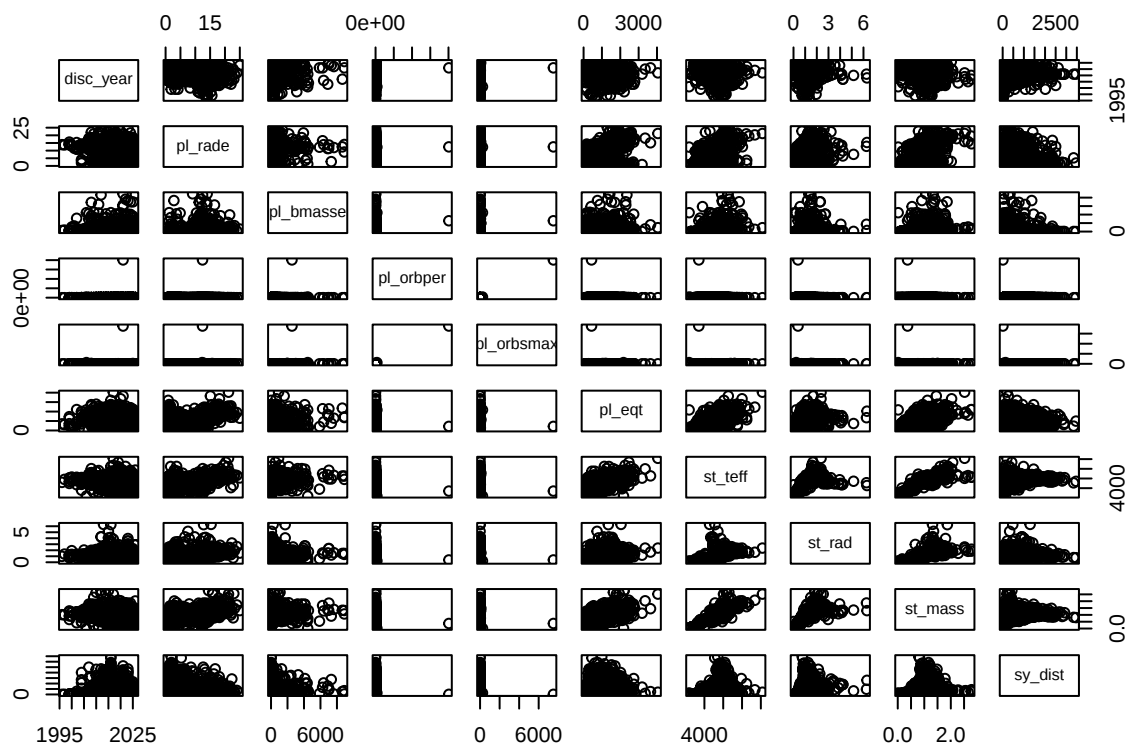
La distribución de `pl_rad` presenta una clara asimetría positiva, con una concentración de planetas de radio reducido y algunos valores extremos de mayor tamaño. Esto anticipa posibles problemas de normalidad en los residuos del modelo lineal.

El análisis descriptivo muestra una elevada dispersión en varias variables. En particular, `pl_bmasse`, `pl_orbper`, `pl_orbsmax` y `sy_dist` presentan rangos amplios, lo que indica que la muestra incluye exoplanetas y sistemas planetarios muy heterogéneos. Esta variabilidad es esperable en una base de datos astronómica, ya que recoge planetas de características físicas y orbitales muy distintas.

También se observa que algunas variables presentan diferencias importantes entre la media y la mediana, lo que sugiere distribuciones asimétricas y posible presencia de valores extremos. Este aspecto deberá tenerse en cuenta en el ajuste posterior del modelo, ya que puede afectar a la calidad de los residuos y justificar el uso de transformaciones en el apartado de mejora.

2.2 Análisis gráfico

```
plot(datos)
```



El gráfico de dispersión por pares permite observar visualmente las relaciones bivariantes entre las variables seleccionadas. En general, no se aprecian relaciones lineales fuertes entre `pl_rade` y la mayoría de variables explicativas. Sin embargo, sí se observa una relación muy marcada entre `pl_orbper` y `pl_orbsmax`, lo que anticipa un posible problema de multicolinealidad entre ambas variables orbitales.

2.3 Correlaciones con la variable dependiente

```
# Correlaciones con la variable dependiente
cor(datos)["pl_rade", ]
```

```
##      disc_year      pl_rade      pl_bmasse      pl_orbper      pl_orbsmax
## -0.0008999011  1.0000000000  0.4240659291  0.0264990613  0.0283654983
##           pl_eqt      st_teff           st_rad           st_mass           sy_dist
##  0.4244222327  0.2730253592  0.4068888520  0.4115688401 -0.1405485372
```

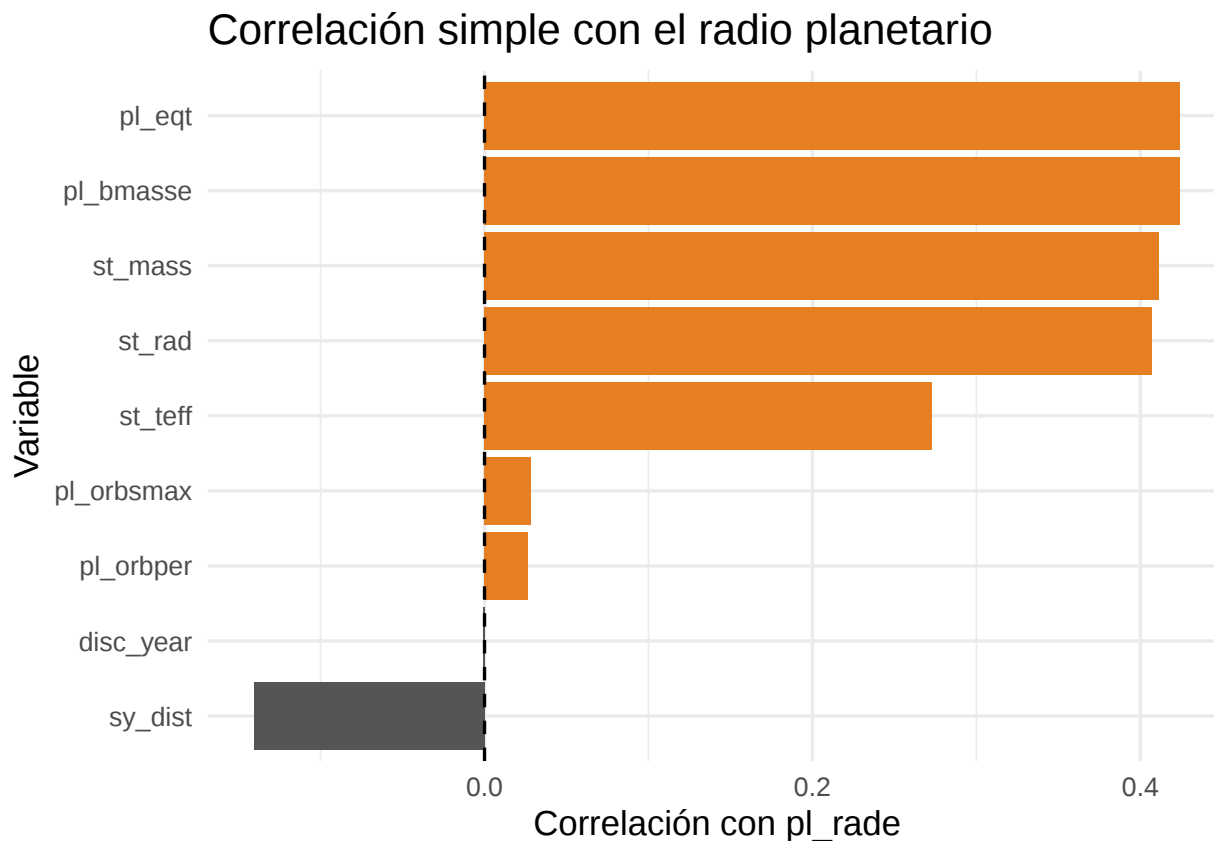
```
matriz_cor_simple <- cor(datos, use = "complete.obs")
round(matriz_cor_simple, 3)
```

```
##           disc_year pl_rade pl_bmasse pl_orbper pl_orbsmax pl_eqt st_teff
## disc_year      1.000  -0.001   0.008   0.013       0.013  -0.056  -0.189
## pl_rade       -0.001   1.000   0.424   0.026       0.028   0.424   0.273
## pl_bmasse      0.008   0.424   1.000   0.067       0.072   0.184   0.147
```

```
## pl_orbper      0.013   0.026   0.067   1.000   1.000 -0.015 -0.037
## pl_orbsmax     0.013   0.028   0.072   1.000   1.000 -0.015 -0.036
## pl_eqt        -0.056   0.424   0.184  -0.015  -0.015  1.000  0.459
## st_teff       -0.189   0.273   0.147  -0.037  -0.036  0.459  1.000
## st_rad        -0.086   0.407   0.201  -0.022  -0.022  0.463  0.683
## st_mass       -0.151   0.412   0.208  -0.032  -0.031  0.522  0.897
## sy_dist       -0.258  -0.141  -0.080  -0.019  -0.019  0.037  0.438
##              st_rad st_mass sy_dist
## disc_year    -0.086  -0.151  -0.258
## pl_rade       0.407   0.412  -0.141
## pl_bmasse     0.201   0.208  -0.080
## pl_orbper    -0.022  -0.032  -0.019
## pl_orbsmax   -0.022  -0.031  -0.019
## pl_eqt       0.463   0.522  0.037
## st_teff      0.683   0.897  0.438
## st_rad       1.000   0.817  0.241
## st_mass      0.817   1.000  0.322
## sy_dist      0.241   0.322  1.000
```

```
cor_pl_rade <- cor(datos)["pl_rade", ] %>%
  as.data.frame() %>%
  rownames_to_column("variable") %>%
  rename(correlacion = 2) %>%
  filter(variable != "pl_rade") %>%
  mutate(variable = reorder(variable, correlacion))

ggplot(cor_pl_rade, aes(x = correlacion, y = variable, fill = correlacion > 0)) +
  geom_col(show.legend = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  scale_fill_manual(values = c("#555555", "#E67E22")) +
  labs(
    title = "Correlación simple con el radio planetario",
    x = "Correlación con pl_rade",
    y = "Variable"
  ) +
  theme_minimal(base_size = 13)
```



La matriz de correlaciones simples muestra la relación lineal entre cada variable numérica y la variable dependiente `pl_rade`. Como era esperable, la correlación de `pl_rade` consigo misma es igual a 1, por lo que no se interpreta.

Las variables que presentan una mayor asociación lineal con el radio planetario son `pl_eqt` y `pl_bmasse`, ambas con una correlación aproximada de 0.424. Esto indica una relación positiva moderada: los planetas con mayor temperatura de equilibrio o mayor masa tienden, en promedio, a presentar radios superiores.

También se observan correlaciones positivas moderadas con `st_mass` y `st_rad`, con valores cercanos a 0.41. Esto sugiere que las características de la estrella anfitriona, especialmente su masa y su radio, podrían estar relacionadas con el tamaño del planeta.

La variable `st_teff` presenta una correlación positiva más débil, de aproximadamente 0.273, mientras que `sy_dist` muestra una correlación negativa débil, cercana a -0.141. Por tanto, su relación lineal directa con el radio planetario parece limitada.

Por último, `pl_orbper`, `pl_orbsmax` y `disc_year` presentan correlaciones prácticamente nulas con `pl_rade`. Esto indica que, de forma simple, no muestran una relación lineal apreciable con el radio del planeta. No obstante, `pl_orbper` y `pl_orbsmax` se analizarán con más detalle posteriormente, ya que pueden presentar multicolinealidad entre ellas.

2.4 Análisis previo de multicolinealidad

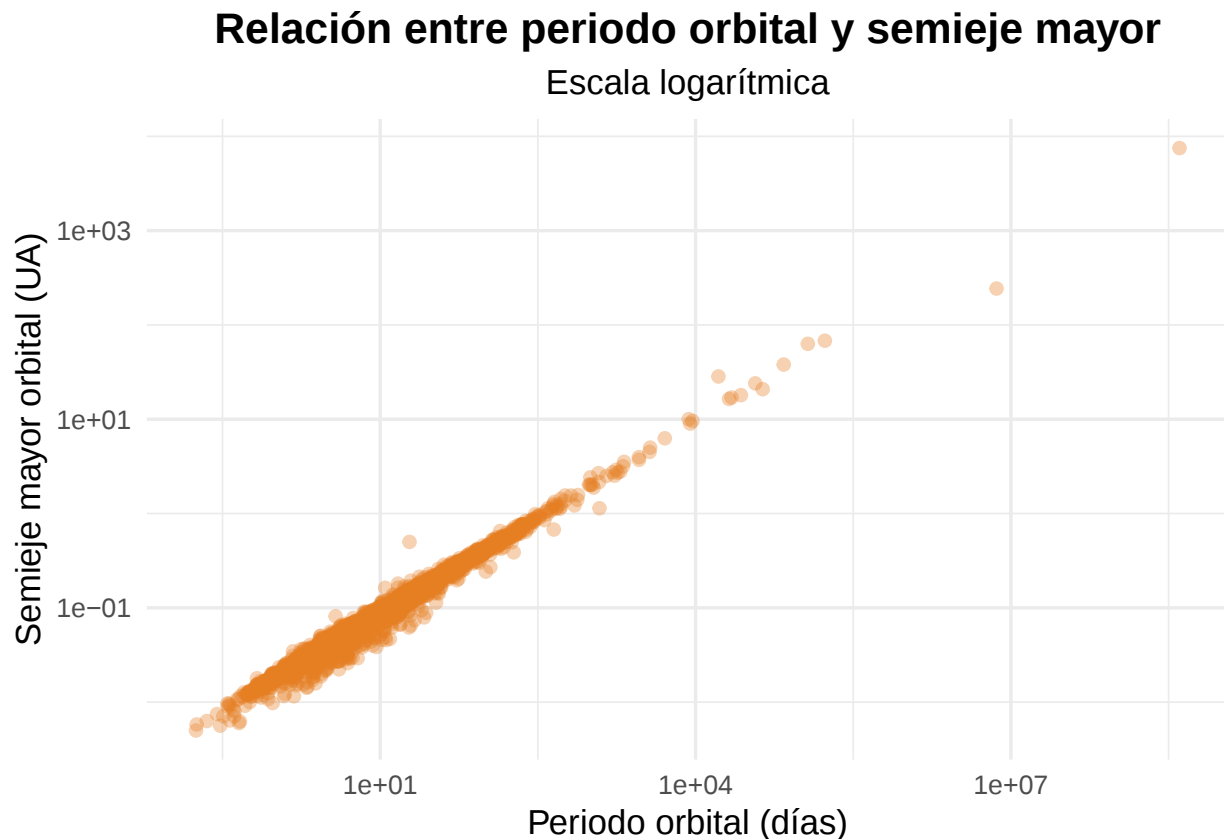
Antes de calcular la matriz de correlaciones parciales, se revisa la matriz de correlaciones simples entre las variables explicativas. Este paso es importante porque las correlaciones parciales requieren invertir una matriz de covarianzas/correlaciones, y si existen variables casi perfectamente correlacionadas entre sí pueden aparecer problemas numéricos.

En particular, se analiza la relación entre `pl_orbper` y `pl_orbsmax`, ya que ambas variables describen aspectos orbitales relacionados: el periodo orbital y el semieje mayor de la órbita.

```
cor(datos$pl_orbper, datos$pl_orbsmax)
```

```
## [1] 0.9997907
```

```
ggplot(datos, aes(x = pl_orbper, y = pl_orbsmax)) +  
  geom_point(alpha = 0.35, color = "#E67E22") +  
  scale_x_log10() +  
  scale_y_log10() +  
  labs(  
    title = "Relación entre periodo orbital y semieje mayor",  
    subtitle = "Escala logarítmica",  
    x = "Periodo orbital (días)",  
    y = "Semieje mayor orbital (UA)"  
  ) +  
  theme_minimal(base_size = 13) +  
  theme(  
    plot.title = element_text(hjust = 0.5, face = "bold"),  
    plot.subtitle = element_text(hjust = 0.5)  
  )
```



El resultado muestra una correlación extremadamente elevada entre `pl_orbper` y `pl_orbsmax`, cercana a 1. Esto indica que ambas variables aportan información muy similar. Desde un punto de vista físico, este

resultado es coherente, ya que los planetas más alejados de su estrella suelen tener periodos orbitales más largos.

Para comprobar cómo afecta esta relación al cálculo de correlaciones parciales, se intenta calcular inicialmente la matriz con todas las variables seleccionadas:

```
# Intento inicial de calcular correlaciones parciales con todas las variables
pcor(datos)$estimate
```

```
## Error in solve.default(cvx): sistema es computacionalmente singular: número de condición recíproco =
```

El cálculo anterior puede generar un problema numérico debido a la elevada multicolinealidad entre `pl_orbper` y `pl_orbsmax`. En concreto, la función `pcor()` necesita invertir internamente una matriz, y esta operación puede fallar cuando dos variables están casi perfectamente correlacionadas.

Por este motivo, se decide no incluir ambas variables simultáneamente. Siguiendo el criterio de conservar la variable que presenta mayor correlación lineal con la variable dependiente `pl_rade`, se comparan ambas correlaciones:

```
# Comparación de la relación de cada variable orbital con la variable dependiente
cor(datos$pl_rade, datos$pl_orbper)
```

```
## [1] 0.02649906
```

```
cor(datos$pl_rade, datos$pl_orbsmax)
```

```
## [1] 0.0283655
```

Como `pl_orbsmax` presenta una correlación ligeramente superior con `pl_rade`, se conserva esta variable y se elimina `pl_orbper` para evitar problemas de multicolinealidad.

```
datos <- datos %>%
  dplyr::select(
    pl_rade,
    pl_bmasse,
    pl_orbsmax,
    pl_eqt,
    st_teff,
    st_rad,
    st_mass,
    sy_dist
  )
```

2.5 Correlaciones parciales con la variable dependiente

```
# Correlaciones parciales con la variable dependiente
pcor(datos)$estimate["pl_rade", ]
```

```
##      pl_rade  pl_bmasse  pl_orbsmax    pl_eqt    st_teff    st_rad
## 1.00000000  0.34081003  0.01777395  0.22695490 -0.13176927  0.07968414
##      st_mass    sy_dist
## 0.18257761 -0.19586252
```

```
matriz_cor_parcial <- pcor(datos)$estimate
round(matriz_cor_parcial, 3)
```

```
##          pl_rade pl_bmasse pl_orbsmax pl_eqt st_teff st_rad st_mass sy_dist
## pl_rade      1.000      0.341      0.018  0.227 -0.132  0.080   0.183  -0.196
## pl_bmasse     0.341      1.000      0.068 -0.018   0.015  0.007   0.013  -0.048
## pl_orbsmax    0.018      0.068      1.000 -0.009 -0.009 -0.001  -0.011   0.010
## pl_eqt        0.227     -0.018     -0.009  1.000   0.079  0.053   0.101  -0.106
## st_teff       -0.132     0.015     -0.009  0.079   1.000 -0.186   0.786   0.316
## st_rad        0.080     0.007     -0.001  0.053  -0.186  1.000   0.578   0.063
## st_mass       0.183     0.013     -0.011  0.101   0.786  0.578   1.000  -0.077
## sy_dist      -0.196     -0.048      0.010 -0.106   0.316  0.063  -0.077   1.000
```

Las correlaciones parciales muestran que, al controlar por el resto de variables, `pl_bmasse` sigue siendo la variable más relacionada con `pl_rade`, con una correlación parcial positiva de 0.341. También mantienen cierta asociación `pl_eqt`, `sy_dist` y `st_mass`, aunque con valores más moderados. En cambio, `pl_orbsmax` y `st_rad` presentan correlaciones parciales muy bajas, lo que sugiere que su aportación específica al explicar el radio planetario es limitada.

Además, `st_teff` cambia de signo respecto a la correlación simple, pasando de una relación positiva a una relación parcial negativa. Esto sugiere que la asociación simple entre la temperatura de la estrella y el radio planetario podía estar condicionada por su relación con otras variables del modelo.

2.6 Conclusión del análisis exploratorio

A partir del análisis exploratorio se observa que las variables con mayor relación lineal simple con `pl_rade` son `pl_bmasse`, `pl_eqt`, `st_mass` y `st_rad`. No obstante, al analizar las correlaciones parciales, la variable que mantiene una asociación más clara con el radio planetario es `pl_bmasse`, seguida de `pl_eqt`, `sy_dist` y `st_mass`.

También se detecta una correlación extremadamente elevada entre `pl_orbper` y `pl_orbsmax`, lo que indica un problema de multicolinealidad. Por este motivo, no se incluirán ambas variables simultáneamente en el modelo. Dado que `pl_orbsmax` presenta una correlación ligeramente superior con `pl_rade`, se conserva esta variable y se descarta `pl_orbper`.

Por tanto, para el ajuste inicial del modelo de regresión múltiple se considerarán como variables explicativas `pl_bmasse`, `pl_orbsmax`, `pl_eqt`, `st_teff`, `st_rad`, `st_mass` y `sy_dist`.

3 Ajuste del modelo lineal

```
fit <- lm(pl_rade ~ ., data = datos)

sfit <- summary(fit)
sfit

##
## Call:
## lm(formula = pl_rade ~ ., data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9195  -2.1529  -0.6519   1.0648  15.4642
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7440947   0.4862626   5.643 1.77e-08 ***
## pl_bmasse    0.0024233   0.0001013  23.931 < 2e-16 ***
## pl_orbsmax   0.0005660   0.0004823   1.174   0.241
## pl_eqt       0.0021959   0.0001427  15.384 < 2e-16 ***
## st_teff     -0.0014160   0.0001614  -8.775 < 2e-16 ***
## st_rad       1.1795106   0.2235132   5.277 1.38e-07 ***
## st_mass      7.3293325   0.5978764  12.259 < 2e-16 ***
## sy_dist     -0.0017497   0.0001327 -13.185 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.608 on 4358 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3952
## F-statistic: 408.4 on 7 and 4358 DF,  p-value: < 2.2e-16
```

El modelo lineal múltiple se ajusta utilizando `pl_rade` como variable dependiente y como variables explicativas `pl_bmasse`, `pl_orbsmax`, `pl_eqt`, `st_teff`, `st_rad`, `st_mass` y `sy_dist`. La variable `pl_orbper` no se incluye debido a la multicolinealidad detectada previamente con `pl_orbsmax`.

El modelo global resulta estadísticamente significativo, ya que el contraste F presenta un p-valor inferior a 0.001. Por tanto, se rechaza la hipótesis nula de que todos los coeficientes del modelo sean simultáneamente iguales a cero. Esto indica que, en conjunto, las variables explicativas incluidas aportan información significativa para explicar el radio planetario.

El coeficiente de determinación múltiple es $R^2 = 0.3962$, mientras que el R^2 ajustado es 0.3952. Esto significa que el modelo explica aproximadamente el 39.5 % de la variabilidad observada en el radio de los exoplanetas. Aunque el modelo tiene capacidad explicativa, queda una parte importante de la variabilidad sin explicar, lo cual es razonable dada la complejidad física de los sistemas planetarios y la diversidad de tipos de exoplanetas incluidos en la muestra.

```
# Tabla de coeficientes
sfit$coefficients
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 2.7440946621 0.4862626062  5.643236 1.774673e-08
## pl_bmasse    0.0024233337 0.0001012619 23.931342 5.085346e-119
```

```
## pl_orbsmax    0.0005660246 0.0004823243    1.173535 2.406453e-01
## pl_eqt        0.0021959240 0.0001427417   15.383900 4.790326e-52
## st_teff       -0.0014159586 0.0001613576   -8.775285 2.406850e-18
## st_rad        1.1795105767 0.2235131926    5.277141 1.375900e-07
## st_mass       7.3293325476 0.5978763670   12.258943 5.429768e-34
## sy_dist       -0.0017496880 0.0001327002  -13.185274 5.889055e-39
```

En cuanto a los coeficientes individuales, la mayoría de variables resultan estadísticamente significativas al nivel del 5 %. En concreto, `pl_bmasse`, `pl_eqt`, `st_teff`, `st_rad`, `st_mass` y `sy_dist` presentan p-valores inferiores a 0.001, por lo que existe evidencia estadística de que están asociadas con `pl_rade`, manteniendo constantes el resto de variables.

La única variable no significativa es `pl_orbsmax`, cuyo p-valor es aproximadamente 0.241. Esto indica que, una vez controlado el efecto del resto de variables, no hay evidencia suficiente para afirmar que el semieje mayor orbital tenga un efecto lineal significativo sobre el radio planetario.

El intercepto del modelo es igual a 2.7441 radios terrestres. Este valor representa el radio planetario esperado cuando todas las variables explicativas del modelo toman simultáneamente el valor cero, es decir, un planeta con masa nula, semieje mayor orbital nulo, temperatura de equilibrio nula, temperatura estelar nula, radio estelar nulo, masa estelar nula y distancia al sistema nula. Esta situación carece de interpretación física real, ya que ningún exoplaneta puede presentar simultáneamente esas condiciones. Por tanto, el intercepto no debe interpretarse de forma sustantiva en este contexto, sino que forma parte de la especificación matemática del modelo y garantiza que la recta de regresión esté correctamente posicionada en el espacio de las variables.

```
# Intervalos de confianza para los coeficientes
confint(fit)
```

```
##                2.5 %        97.5 %
## (Intercept)  1.7907726983  3.697416626
## pl_bmasse    0.0022248089  0.002621859
## pl_orbsmax   -0.0003795762  0.001511626
## pl_eqt       0.0019160777  0.002475770
## st_teff      -0.0017323014 -0.001099616
## st_rad       0.7413110668  1.617710087
## st_mass      6.1571908591  8.501474236
## sy_dist      -0.0020098478 -0.001489528
```

```
# Intervalos de confianza de los coeficientes
coef_ci <- tidy(fit, conf.int = TRUE) %>%
  filter(term != "(Intercept)") %>%
  mutate(
    significativo = ifelse(conf.low > 0 | conf.high < 0,
                          "Significativo",
                          "No significativo"),
    term = reorder(term, estimate)
  )

ggplot(coef_ci, aes(x = estimate, y = term, color = significativo)) +
  geom_vline(xintercept = 0, linetype = "dashed", linewidth = 0.8, color = "black") +
  geom_errorbarh(
    aes(xmin = conf.low, xmax = conf.high),
    height = 0.22,
    linewidth = 0.9
  ) +
```

```

geom_point(size = 3.5) +
scale_x_continuous(
trans = scales::pseudo_log_trans(base = 10),
labels = scales::label_number(accuracy = 0.001)
) +
scale_color_manual(
values = c(
"Significativo" = "#E67E22",
"No significativo" = "#555555"
)
) +
labs(
title = "Intervalos de confianza al 95% de los coeficientes",
x = "Estimación del coeficiente",
y = "Variable explicativa",
color = ""
) +
theme_minimal(base_size = 13) +
theme(
plot.title = element_text(face = "bold", size = 17),
plot.subtitle = element_text(size = 12),
axis.title = element_text(face = "bold"),
legend.position = "top",
panel.grid.minor = element_blank()
)

```

```

## Warning: 'geom_errorbarh()' was deprecated in ggplot2 4.0.0.
## i Please use the 'orientation' argument of 'geom_errorbar()' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.

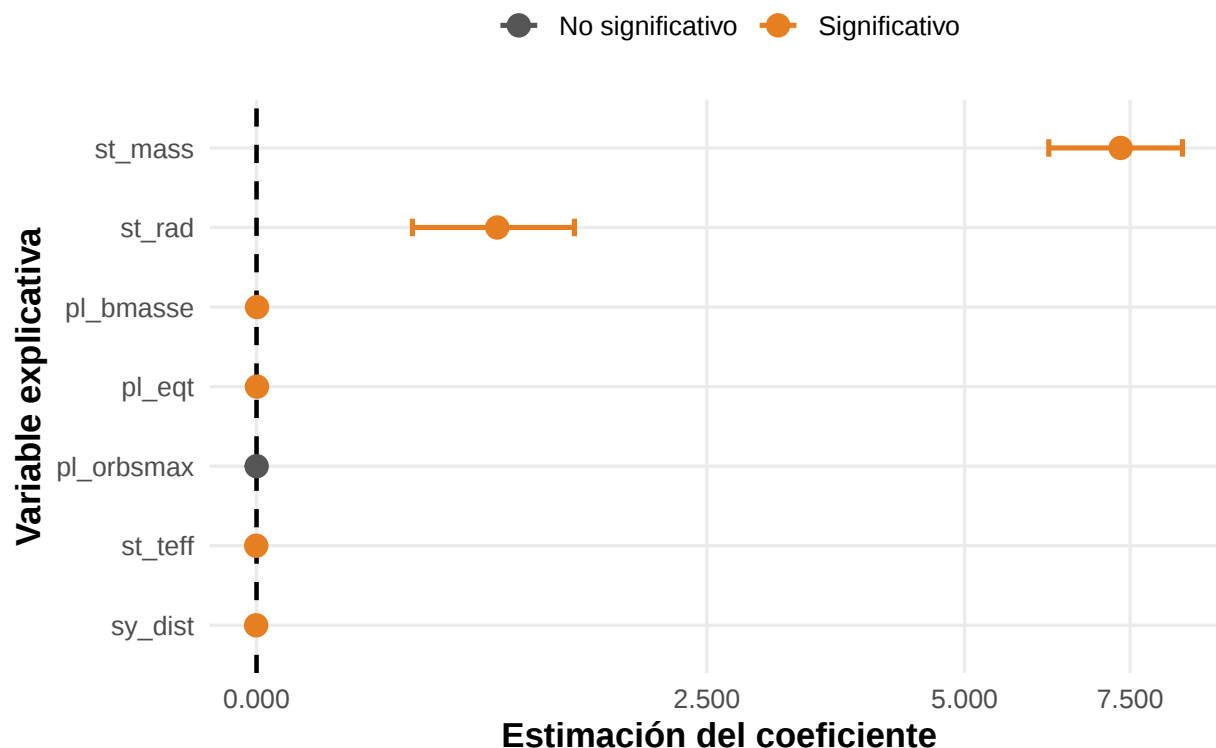
```

```

## 'height' was translated to 'width'.

```

Intervalos de confianza al 95% de los coeficientes



El coeficiente de **pl_bmasse** es positivo e igual a 0.00242. Esto significa que, manteniendo constantes el resto de variables, un aumento de una masa terrestre en la masa del planeta se asocia con un incremento esperado de aproximadamente 0.0024 radios terrestres en el radio planetario. Aunque el efecto unitario es pequeño, la variable es altamente significativa.

El coeficiente de **pl_eqt** también es positivo, con un valor aproximado de 0.00220. Esto indica que, manteniendo constantes las demás variables, un aumento de un Kelvin en la temperatura de equilibrio se asocia con un incremento esperado de 0.0022 radios terrestres en el radio planetario. En términos prácticos, un aumento de 100 K se asociaría con un incremento aproximado de 0.22 radios terrestres.

El coeficiente de **st_mass** es positivo y relativamente elevado, con un valor de 7.329. Esto sugiere que, manteniendo constantes el resto de variables, los planetas que orbitan estrellas de mayor masa tienden a presentar radios mayores. En concreto, un aumento de una masa solar en la masa de la estrella se asocia con un incremento esperado de aproximadamente 7.33 radios terrestres en el radio planetario.

El coeficiente de **st_teff** es negativo, con un valor de -0.00142. Aunque en las correlaciones simples la temperatura efectiva de la estrella presentaba una asociación positiva con **pl_rade**, en el modelo múltiple su efecto estimado pasa a ser negativo al controlar por el resto de variables. Esto indica que su relación con el radio planetario está condicionada por otras variables estelares, como **st_mass** y **st_rad**.

La variable **sy_dist** también presenta un coeficiente negativo, de aproximadamente -0.00175. Esto sugiere que, manteniendo constantes el resto de variables, los sistemas más lejanos tienden a asociarse con radios planetarios ligeramente menores en esta muestra. Esta relación puede estar influida por sesgos observacionales o por las características de los planetas mejor caracterizados.

3.1 Conclusión

Desde el punto de vista interpretativo, el modelo sugiere que el radio planetario está especialmente relacionado con la masa del planeta, la temperatura de equilibrio y las características de la estrella anfitriona. En particular, planetas más masivos, con mayor temperatura de equilibrio y orbitando estrellas de mayor masa tienden a presentar radios mayores.

No obstante, estas conclusiones deben interpretarse con cautela. El objetivo no es establecer una relación causal directa, sino identificar asociaciones estadísticas en la muestra de exoplanetas confirmados. Además, el valor de R^2 indica que el modelo explica una parte relevante, pero no mayoritaria, de la variabilidad del radio planetario, por lo que existen otros factores físicos no incorporados al modelo que también pueden influir en el tamaño del planeta, como la composición, la densidad, la atmósfera o la historia evolutiva del sistema.

Para identificar las variables con mayor impacto sobre el radio planetario, se ajusta también un modelo con las variables estandarizadas. Esta transformación permite comparar los coeficientes entre sí, ya que todas las variables pasan a estar medidas en la misma escala. De esta forma, el tamaño absoluto del coeficiente estandarizado refleja la importancia relativa de cada variable en el modelo.

```
# Modelo con variables estandarizadas
datos_estandarizados <- datos %>%
  mutate(across(everything(), scale))

fit_estandarizado <- lm(pl_rade ~ ., data = datos_estandarizados)

# Coeficientes estandarizados
coef_estandarizados <- summary(fit_estandarizado)$coefficients

coef_estandarizados
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	2.635064e-16	0.01176983	2.238830e-14	1.000000e+00
##	pl_bmasse	2.937203e-01	0.01227346	2.393134e+01	5.085346e-119
##	pl_orbsmax	1.386645e-02	0.01181596	1.173535e+00	2.406453e-01
##	pl_eqt	2.163914e-01	0.01406609	1.538390e+01	4.790326e-52
##	st_teff	-2.559607e-01	0.02916836	-8.775285e+00	2.406850e-18
##	st_rad	1.104647e-01	0.02093268	5.277141e+00	1.375900e-07
##	st_mass	4.350909e-01	0.03549171	1.225894e+01	5.429768e-34
##	sy_dist	-1.795005e-01	0.01361371	-1.318527e+01	5.889055e-39

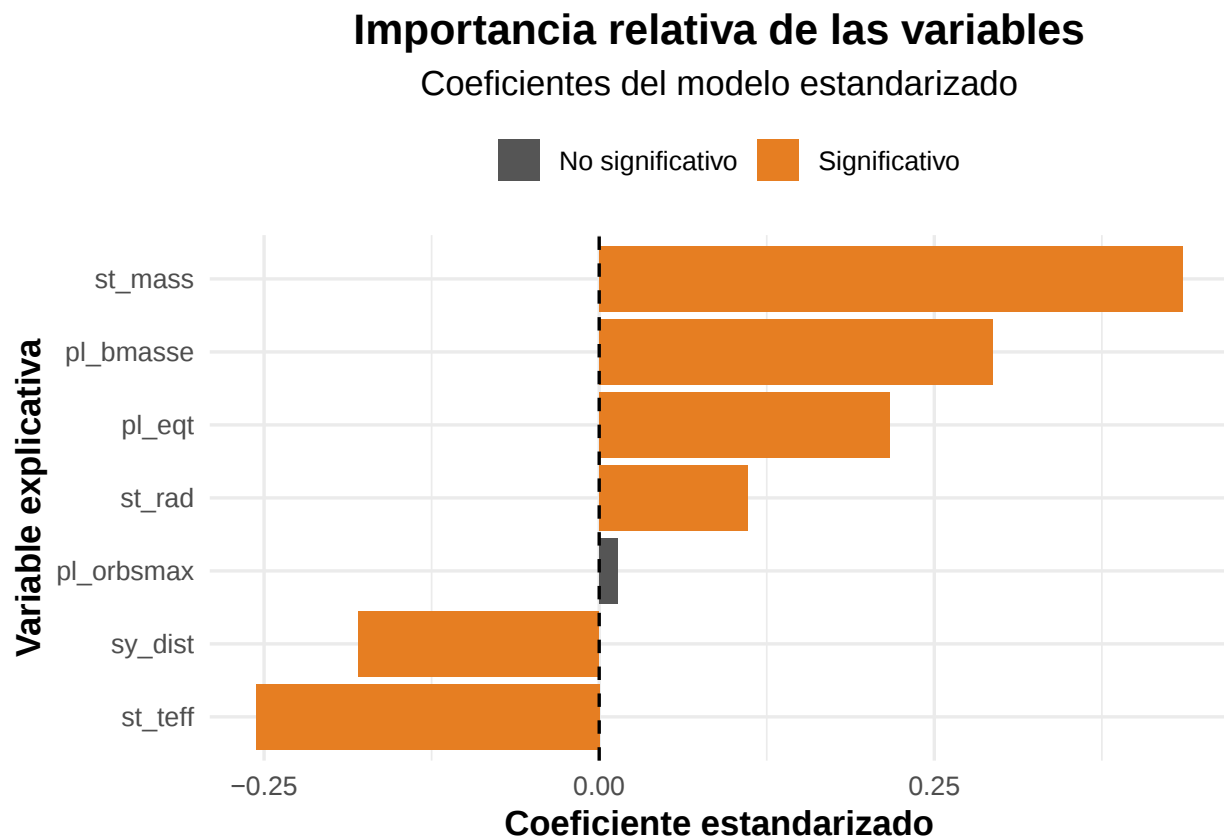
```
coef_std_grafico <- as.data.frame(coef_estandarizados) %>%
  rownames_to_column("variable") %>%
  filter(variable != "(Intercept)") %>%
  mutate(
    significativo = ifelse(`Pr(>|t|)` < 0.05, "Significativo", "No significativo"),
    variable = reorder(variable, Estimate)
  )

ggplot(coef_std_grafico, aes(x = Estimate, y = variable, fill = significativo)) +
  geom_col() +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
  scale_fill_manual(
    values = c(
      "Significativo" = "#E67E22",
```

```

    "No significativo" = "#555555"
  )
) +
labs(
  title = "Importancia relativa de las variables",
  subtitle = "Coeficientes del modelo estandarizado",
  x = "Coeficiente estandarizado",
  y = "Variable explicativa",
  fill = ""
) +
theme_minimal(base_size = 13) +
theme(
  plot.title = element_text(hjust = 0.5, face = "bold"),
  plot.subtitle = element_text(hjust = 0.5),
  axis.title = element_text(face = "bold"),
  legend.position = "top"
)

```



```

# Extraemos los coeficientes sin el intercepto
coef_impacto <- coef_estandarizados[-1, , drop = FALSE]

# Los ordenamos por impacto absoluto
coef_impacto_ordenados <- coef_impacto[order(abs(coef_impacto[, "Estimate"]), decreasing = TRUE), ]

# Mostramos los 3 coeficientes con mayor impacto

```

```
coef_impacto_ordenados[1:3, ]
```

##		Estimate	Std. Error	t value	Pr(> t)
##	st_mass	0.4350909	0.03549171	12.258943	5.429768e-34
##	pl_bmasse	0.2937203	0.01227346	23.931342	5.085346e-119
##	st_teff	-0.2559607	0.02916836	-8.775285	2.406850e-18

Según los coeficientes estandarizados, las tres variables con mayor impacto relativo sobre `pl_rade` son `st_mass`, `pl_bmasse` y `st_teff`.

La variable con mayor efecto es `st_mass`, con un coeficiente positivo de 0.435, lo que indica que los planetas que orbitan estrellas de mayor masa tienden a presentar radios mayores, manteniendo constantes el resto de variables. En segundo lugar aparece `pl_bmasse`, también con efecto positivo, por lo que los planetas más masivos tienden a tener mayor radio. Por último, `st_teff` presenta un coeficiente negativo, lo que sugiere que, al controlar por el resto de variables del modelo, una mayor temperatura efectiva de la estrella se asocia con un menor radio planetario.

Las tres variables son estadísticamente significativas, ya que sus p-valores son inferiores a 0.001.

3.2 Recomendaciones

A partir de los resultados del modelo, se formulan las siguientes recomendaciones orientadas a identificar qué tipo de sistemas planetarios presentan, en promedio, mayores radios planetarios. Estas recomendaciones se basan en el signo, la magnitud, la precisión y la significatividad estadística de los coeficientes estimados, manteniendo siempre constantes el resto de variables del modelo.

Priorizar sistemas con estrellas de mayor masa (`st_mass`). Esta es la variable con mayor impacto relativo según los coeficientes estandarizados (0.435). Su coeficiente es positivo, significativo ($p < 0.001$) e interpretable: un aumento de una masa solar se asocia con un incremento esperado de aproximadamente 7.33 radios terrestres en el radio planetario. El intervalo de confianza al 95 % no incluye el cero, lo que confirma la precisión de esta estimación. En consecuencia, se recomienda **aumentar** esta variable, es decir, orientar la búsqueda hacia planetas que orbiten estrellas más masivas.

Seleccionar planetas con mayor masa (`pl_bmasse`). Con un coeficiente estandarizado de 0.232 y un coeficiente sin estandarizar de 0.00242 ($p < 0.001$), la masa planetaria presenta una relación positiva y significativa con el radio. Aunque el efecto unitario es pequeño, la amplitud reducida del intervalo de confianza indica una estimación precisa. Se recomienda **aumentar** esta variable, es decir, considerar planetas más masivos como candidatos con mayor radio esperado.

Considerar planetas con mayor temperatura de equilibrio (`pl_eqt`). Su coeficiente es positivo (0.00220) y estadísticamente significativo ($p < 0.001$). Un aumento de 100 K en la temperatura de equilibrio se asocia con aproximadamente 0.22 radios terrestres adicionales. Se recomienda **aumentar** esta variable en la selección de candidatos.

Cautela con `st_teff`, `sy_dist` y `pl_orbsmax`. La temperatura efectiva de la estrella presenta un coeficiente negativo en el modelo múltiple (-0.00142), lo que indica que su efecto real está condicionado por otras variables estelares. La distancia al sistema (`sy_dist`) también muestra un coeficiente negativo y significativo, posiblemente influido por sesgos observacionales. Finalmente, `pl_orbsmax` no resulta estadísticamente significativa ($p \approx 0.241$), por lo que no se puede afirmar que el semieje mayor orbital tenga un efecto real sobre el radio planetario y no se formula ninguna recomendación sobre esta variable.

En conjunto, el modelo sugiere que los sistemas con estrellas masivas y planetas igualmente masivos y calientes son los mejores candidatos para encontrar exoplanetas de mayor tamaño. No obstante, estas recomendaciones deben interpretarse como asociaciones estadísticas y no como relaciones causales, dado que el modelo explica únicamente el 39.5 % de la variabilidad del radio planetario.

4 Bondad del ajuste

```
# R^2  
sfit$r.squared
```

```
## [1] 0.3961529
```

```
# R^2 ajustado  
sfit$adj.r.squared
```

```
## [1] 0.395183
```

```
# Estadístico F y p-valor asociado  
# F-statistic: 408.4 on 7 and 4358 DF, p-value: < 2.2e-16
```

```
# CV-sigma  
cv_sigma <- sigma(fit) / mean(datos$pl_rade) * 100  
cv_sigma
```

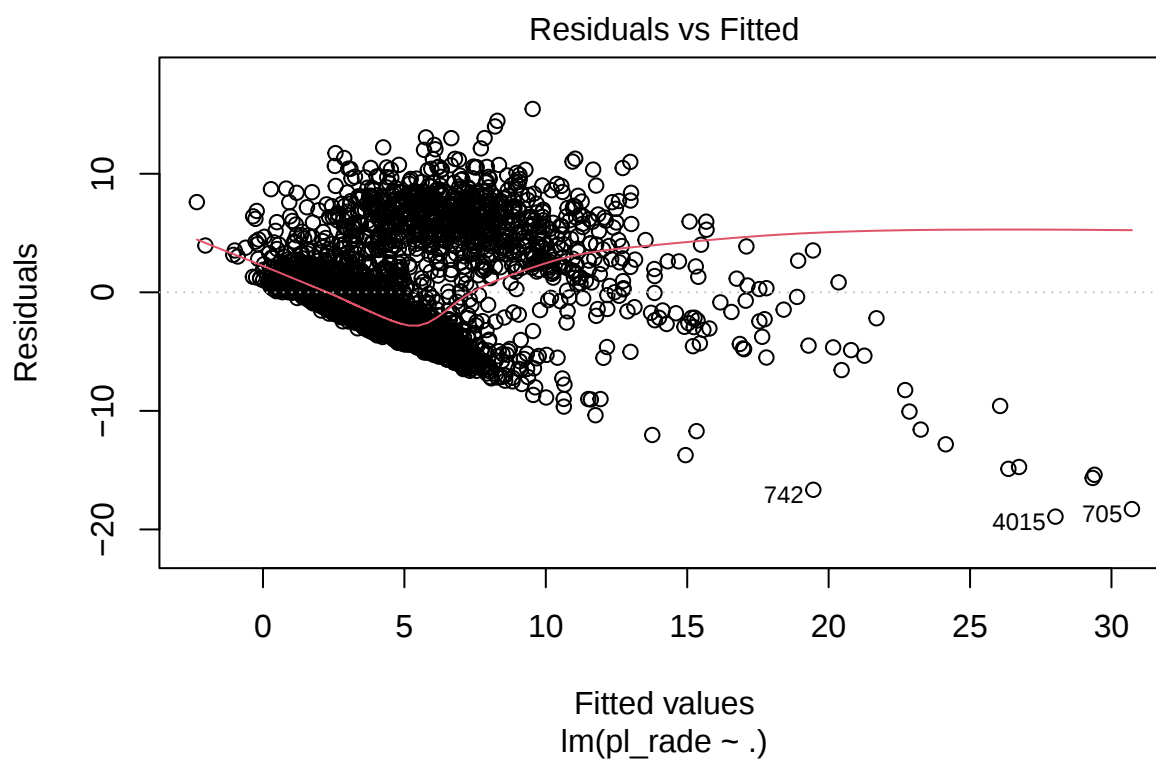
```
## [1] 80.63467
```

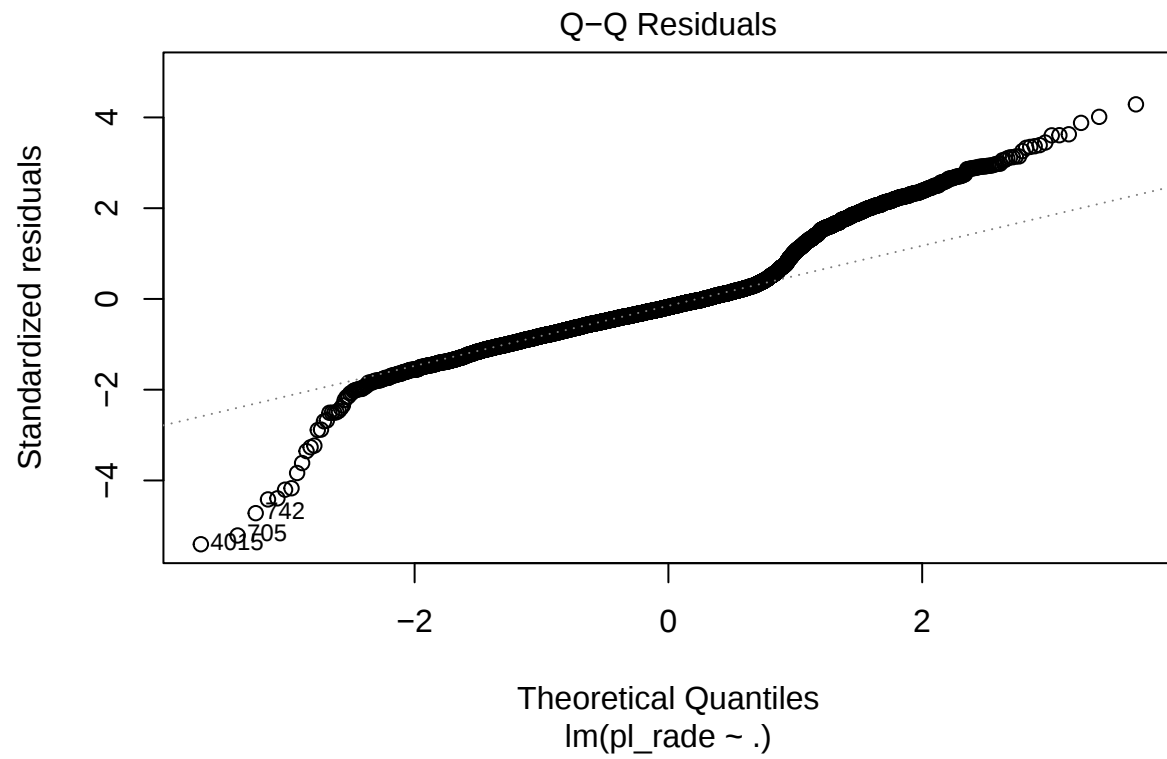
El modelo presenta un R^2 de 0.3962 y un R^2 ajustado de 0.3952, por lo que explica aproximadamente el 39.5 % de la variabilidad del radio planetario. El contraste global del modelo es significativo, ya que el p-valor asociado al estadístico F es inferior a 0.001, lo que indica que las variables explicativas aportan información conjunta sobre `pl_rade`.

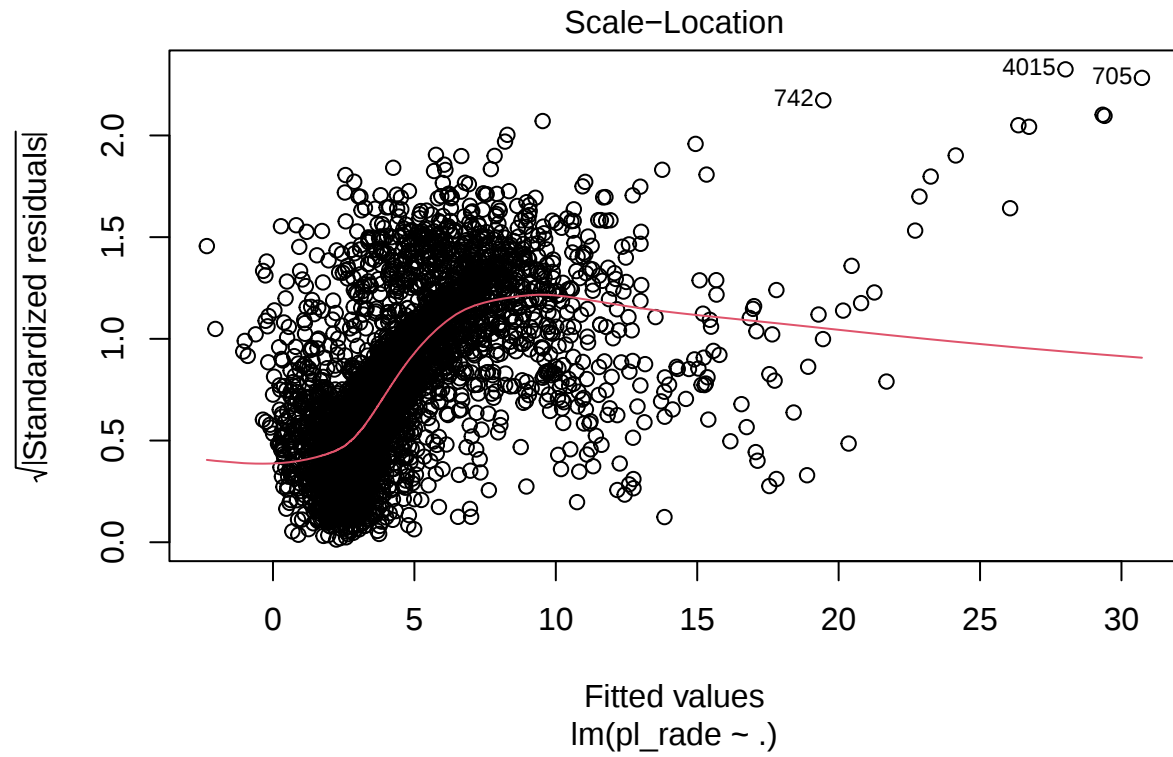
Sin embargo, el valor de CV_σ es elevado, aproximadamente 80.63 %, lo que indica que el error residual es considerable en relación con la media de la variable dependiente. Por tanto, aunque el modelo es estadísticamente significativo, su capacidad predictiva es moderada y podría mejorarse mediante transformaciones o modelos alternativos.

5 Análisis de los residuos

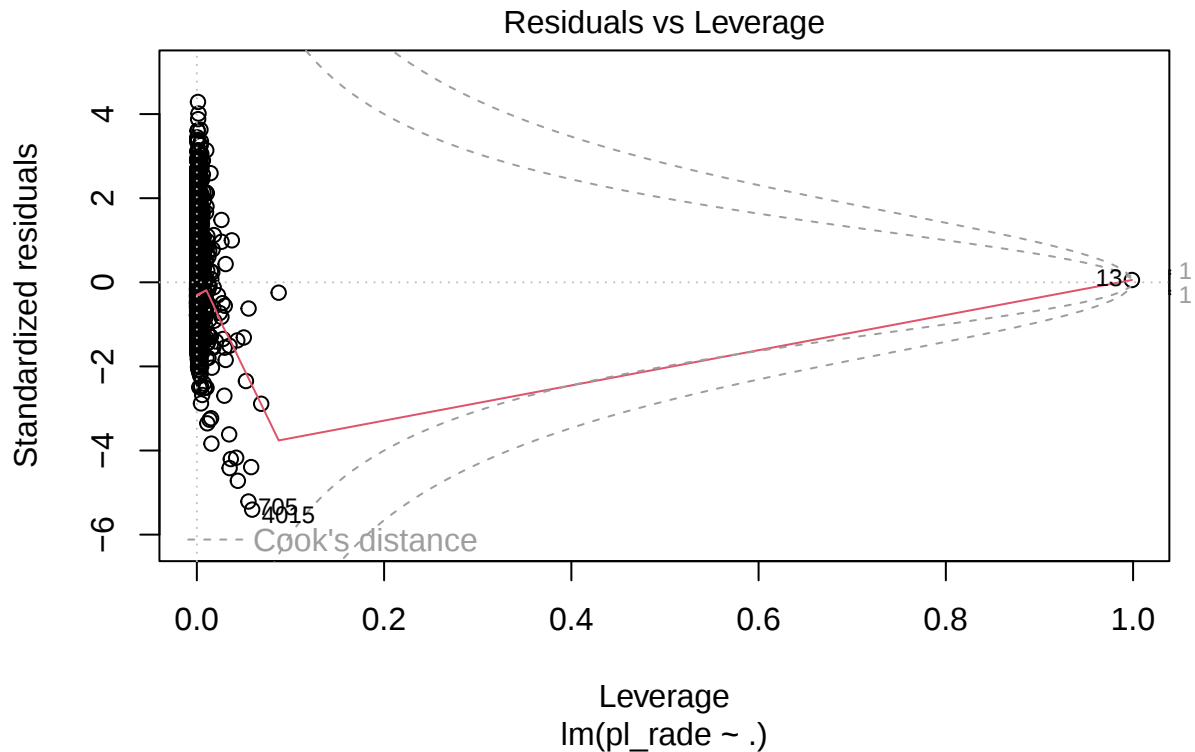
```
plot(fit)
```







```
## Warning in sqrt(crit * p * (1 - hh)/hh): Se han producido NaNs
## Warning in sqrt(crit * p * (1 - hh)/hh): Se han producido NaNs
```

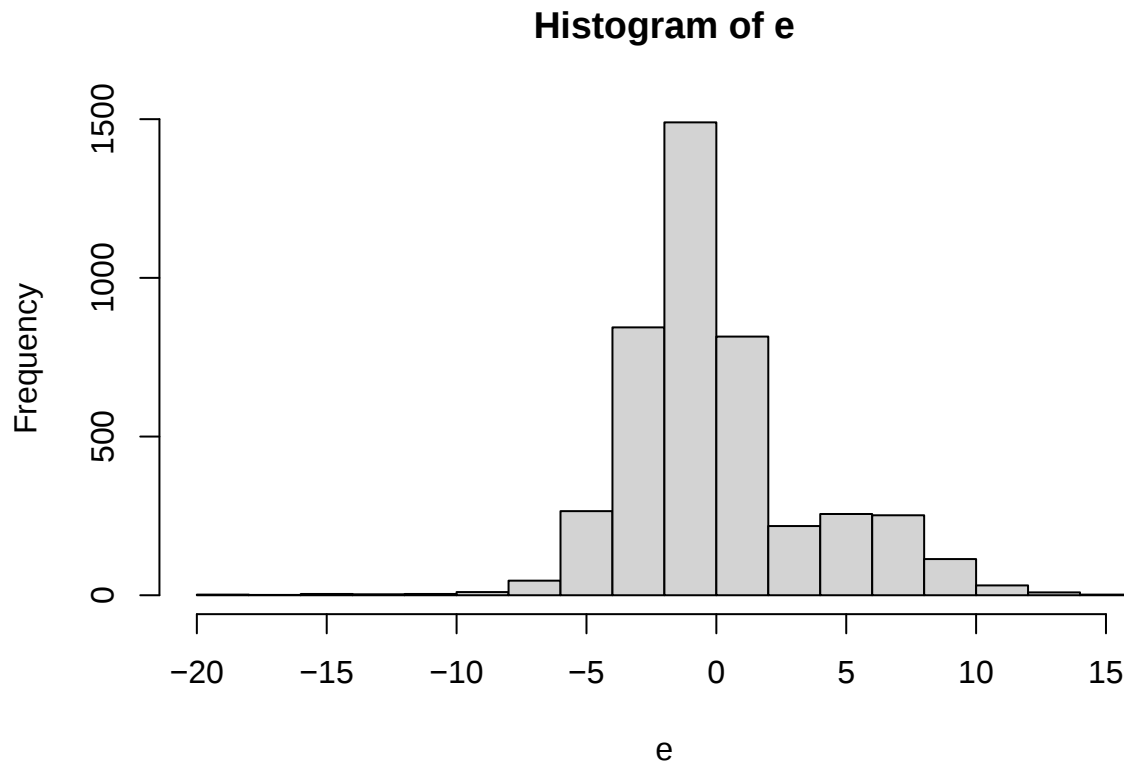


Los gráficos de diagnóstico muestran que el modelo inicial presenta varios problemas. En **Residuals vs Fitted** se observa un patrón no aleatorio, lo que sugiere falta de linealidad y posible heterocedasticidad. El gráfico **Q-Q** muestra desviaciones claras respecto a la normalidad, especialmente en las colas. El gráfico **Scale-Location** confirma que la varianza de los residuos no parece constante. Además, en **Residuals vs Leverage** aparecen algunas observaciones potencialmente influyentes.

Por tanto, el modelo lineal ajustado con las variables originales no cumple completamente los supuestos básicos de regresión. Esto justifica la necesidad de realizar contrastes estadísticos y considerar mejoras posteriores, como transformaciones logarítmicas.

```
# Extraemos residuos
e <- residuals(fit)

# Histograma de residuos
hist(e)
```



El histograma muestra que los residuos se concentran principalmente alrededor de cero, aunque la distribución presenta cierta asimetría y valores extremos en las colas. Por tanto, visualmente no parece cumplirse completamente el supuesto de normalidad de los residuos, algo que se comprobará posteriormente mediante un contraste formal.

```
# Contrastes:
# Contraste de normalidad de Shapiro-Wilk
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.93023, p-value < 2.2e-16
```

El test de Shapiro-Wilk muestra un p-valor inferior a 0.001, por lo que se rechaza la hipótesis nula de normalidad. Esto indica que los residuos no siguen una distribución normal, resultado coherente con el histograma y el gráfico Q-Q de residuos.

```
# Contraste de homocedasticidad de Breusch-Pagan
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data: fit
## BP = 1369.1, df = 7, p-value < 2.2e-16
```

El test de Breusch-Pagan presenta un estadístico $BP = 1369.1$ con 7 grados de libertad y un p-valor inferior a 0.001, por lo que se rechaza la hipótesis nula de homocedasticidad. Esto confirma que la varianza de los residuos no es constante, resultado coherente con el patrón observado en el gráfico Scale-Location. Por tanto, el supuesto de homocedasticidad no se cumple.

```
# Contraste de incorrelación de Durbin-Watson
dwtest(fit, alternative = "two.sided")
```

```
##
## Durbin-Watson test
##
## data: fit
## DW = 1.107, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

El test de Durbin-Watson presenta un estadístico $DW = 1.107$ y un p-valor inferior a 0.001, por lo que se rechaza la hipótesis nula de ausencia de autocorrelación. Esto indica que los residuos presentan autocorrelación, aparentemente positiva al estar el estadístico por debajo de 2. Por tanto, el supuesto de independencia de los errores no se cumple.

5.1 Valores sobreinfluyentes

Para detectar observaciones sobreinfluyentes se utiliza la **distancia de Cook**, que mide cuánto cambian conjuntamente todos los coeficientes estimados si se elimina una observación concreta. Se considera que una observación es sobreinfluyente si su distancia de Cook supera el umbral $4/n$, donde n es el número de observaciones.

```
# Distancia de Cook
cook <- cooks.distance(fit)
n <- nrow(datos)
umbral <- 4 / n

# Número de observaciones sobreinfluyentes
sum(cook > umbral)
```

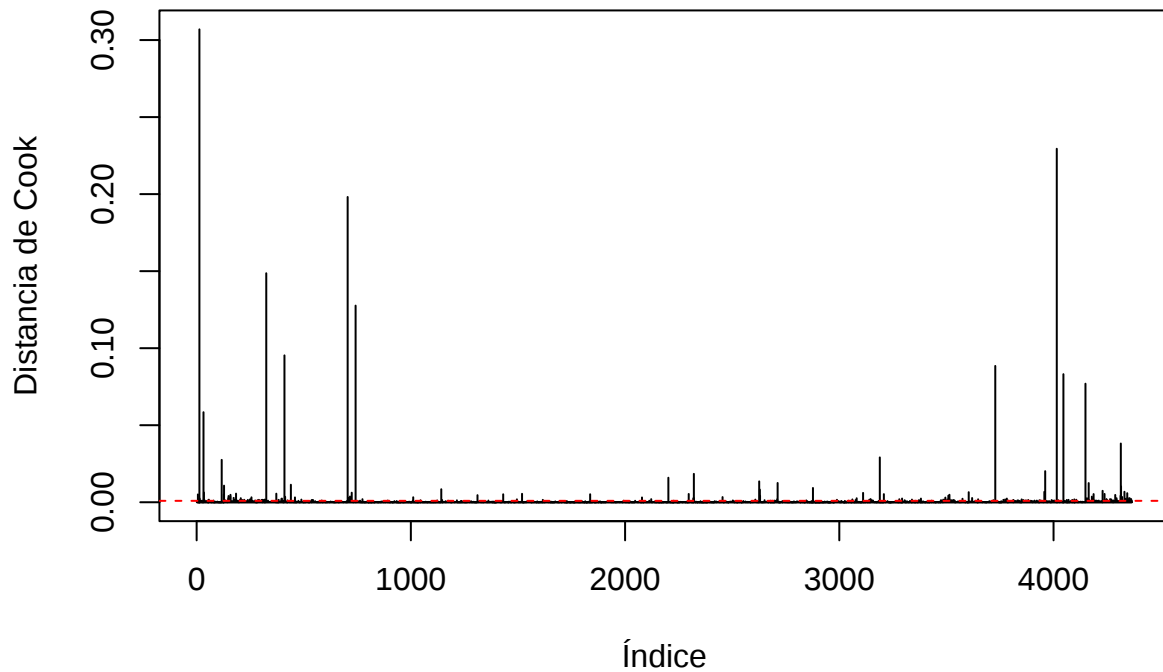
```
## [1] 265
```

```
# Proporción sobre el total
mean(cook > umbral) * 100
```

```
## [1] 6.069629
```

```
# Gráfico de distancias de Cook
plot(cook, type = "h", main = "Distancia de Cook por observación",
      ylab = "Distancia de Cook", xlab = "Índice")
abline(h = umbral, col = "red", lty = 2)
```

Distancia de Cook por observación



El gráfico muestra las distancias de Cook de cada observación junto con el umbral $4/n$ marcado en rojo. Las observaciones que superan dicho umbral se consideran sobreinfluyentes. En este caso, se detectan 265 observaciones sobreinfluyentes, lo que representa aproximadamente el 6.1 % del total de la muestra. Aunque su número es relativamente reducido, su presencia puede estar contribuyendo a los problemas detectados en los supuestos de normalidad y homocedasticidad. Por tanto, el supuesto de ausencia de valores sobreinfluyentes no se cumple de forma completamente satisfactoria.

6 Selección de variables

```
# Selección automática de variables mediante AIC
```

```
fit_step <- step(fit, direction = "backward")
```

```
## Start: AIC=11213.39
```

```
## pl_rade ~ pl_bmasse + pl_orbsmax + pl_eqt + st_teff + st_rad +
```

```
## st_mass + sy_dist
```

```
##
```

```
##           Df Sum of Sq  RSS   AIC
```

```
## - pl_orbsmax  1      17.9 56761 11213
```

```
## <none>                56743 11213
```

```
## - st_rad      1      362.6 57105 11239
```

```
## - st_teff     1     1002.6 57745 11288
```

```
## - st_mass     1     1956.7 58699 11359
```

```
## - sy_dist     1     2263.6 59006 11382
```

```
## - pl_eqt      1     3081.5 59824 11442
```

```
## - pl_bmasse   1     7456.9 64200 11750
```

```
##
```

```
## Step: AIC=11212.77
```

```
## pl_rade ~ pl_bmasse + pl_eqt + st_teff + st_rad + st_mass + sy_dist
```

```
##
```

```
##           Df Sum of Sq  RSS   AIC
```

```
## <none>                56761 11213
```

```
## - st_rad      1      362.7 57123 11239
```

```
## - st_teff     1     1006.0 57767 11288
```

```
## - st_mass     1     1954.0 58715 11358
```

```
## - sy_dist     1     2261.2 59022 11381
```

```
## - pl_eqt      1     3079.4 59840 11441
```

```
## - pl_bmasse   1     7561.8 64322 11757
```

```
# Resumen del modelo seleccionado
```

```
sfit_step <- summary(fit_step)
```

```
sfit_step
```

```
##
```

```
## Call:
```

```
## lm(formula = pl_rade ~ pl_bmasse + pl_eqt + st_teff + st_rad +
```

```
## st_mass + sy_dist, data = datos)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -19.0005 -2.1535 -0.6525  1.0650 15.4632
```

```
##
```

```
## Coefficients:
```

```
##           Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.7608976  0.4860728   5.680 1.43e-08 ***
```

```
## pl_bmasse     0.0024327  0.0001010  24.098 < 2e-16 ***
```

```
## pl_eqt        0.0021952  0.0001427  15.378 < 2e-16 ***
```

```
## st_teff      -0.0014182  0.0001614  -8.790 < 2e-16 ***
```

```
## st_rad        1.1797416  0.2235228   5.278 1.37e-07 ***
```

```
## st_mass       7.3239842  0.5978849  12.250 < 2e-16 ***
```

```
## sy_dist      -0.0017487  0.0001327 -13.178  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.609 on 4359 degrees of freedom
## Multiple R-squared:  0.396, Adjusted R-squared:  0.3951
## F-statistic: 476.2 on 6 and 4359 DF,  p-value: < 2.2e-16
```

Se aplica el procedimiento de selección hacia atrás basado en el criterio AIC. Este algoritmo parte del modelo completo y va eliminando variables cuando su eliminación mejora el AIC. El objetivo es obtener un modelo más simple, manteniendo una capacidad explicativa adecuada.

```
# R2 del modelo simplificado
sfit_step$r.squared
```

```
## [1] 0.3959621
```

```
# R2 ajustado del modelo simplificado
sfit_step$adj.r.squared
```

```
## [1] 0.3951306
```

```
# AIC del modelo completo y del modelo simplificado
AIC(fit, fit_step)
```

```
##          df          AIC
## fit        9 23605.56
## fit_step    8 23604.94
```

```
# CV-sigma del modelo simplificado
cv_sigma_step <- sigma(fit_step) / mean(datos$pl_rade) * 100
cv_sigma_step
```

```
## [1] 80.63816
```

```
comparacion_modelos <- data.frame(
  Modelo = c("Modelo general", "Modelo AIC"),
  R2 = c(
    sfit$r.squared,
    sfit_step$r.squared
  ),
  R2_ajustado = c(
    sfit$adj.r.squared,
    sfit_step$adj.r.squared
  ),
  AIC = c(
    AIC(fit),
    AIC(fit_step)
  ),
  CV_sigma = c(
    sigma(fit) / mean(datos$pl_rade) * 100,
```

```

    sigma(fit_step) / mean(datos$pl_rade) * 100
  )
)

comparacion_modelos

```

```

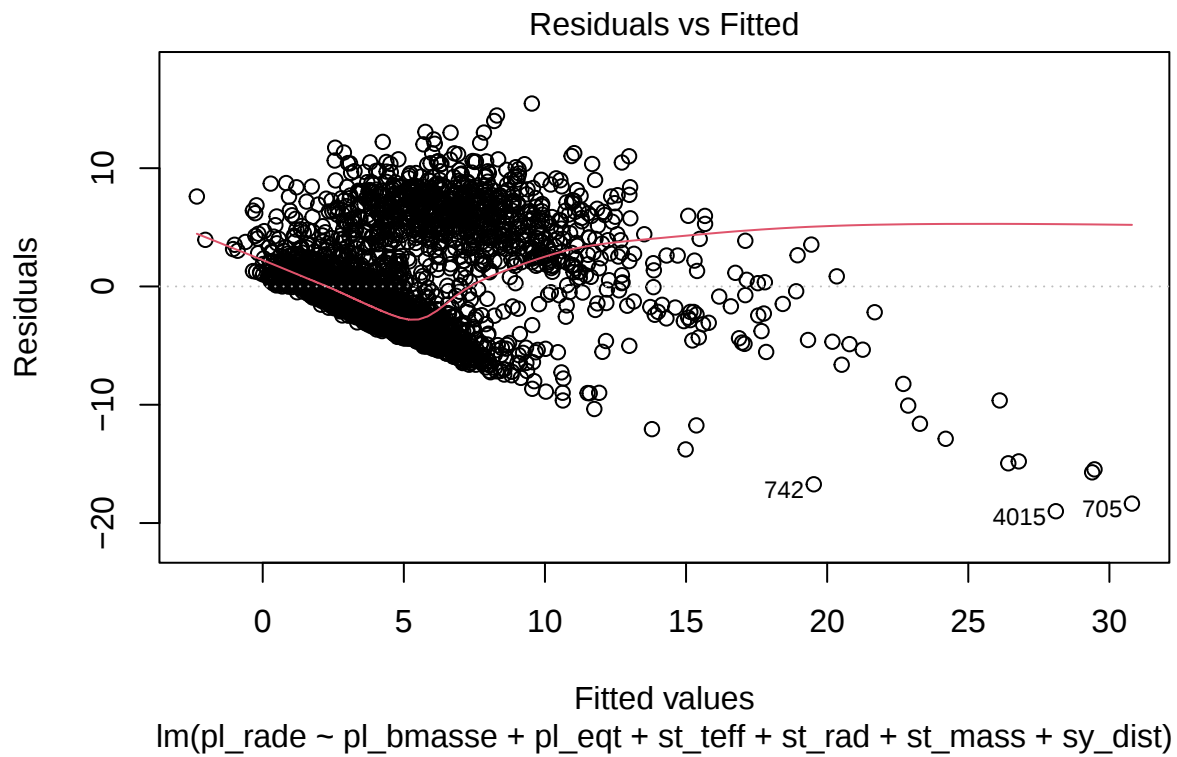
##           Modelo      R2 R2_ajustado      AIC CV_sigma
## 1 Modelo general 0.3961529  0.3951830 23605.56 80.63467
## 2      Modelo AIC 0.3959621  0.3951306 23604.94 80.63816

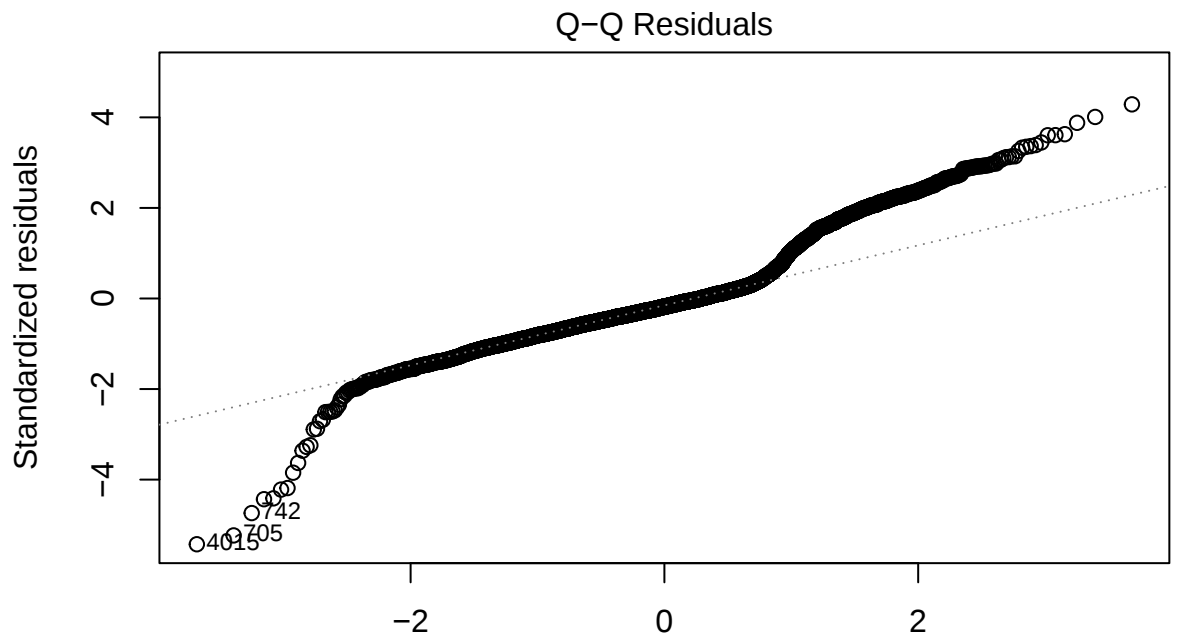
```

```

# Gráficos de diagnóstico del modelo AIC
plot(fit_step)

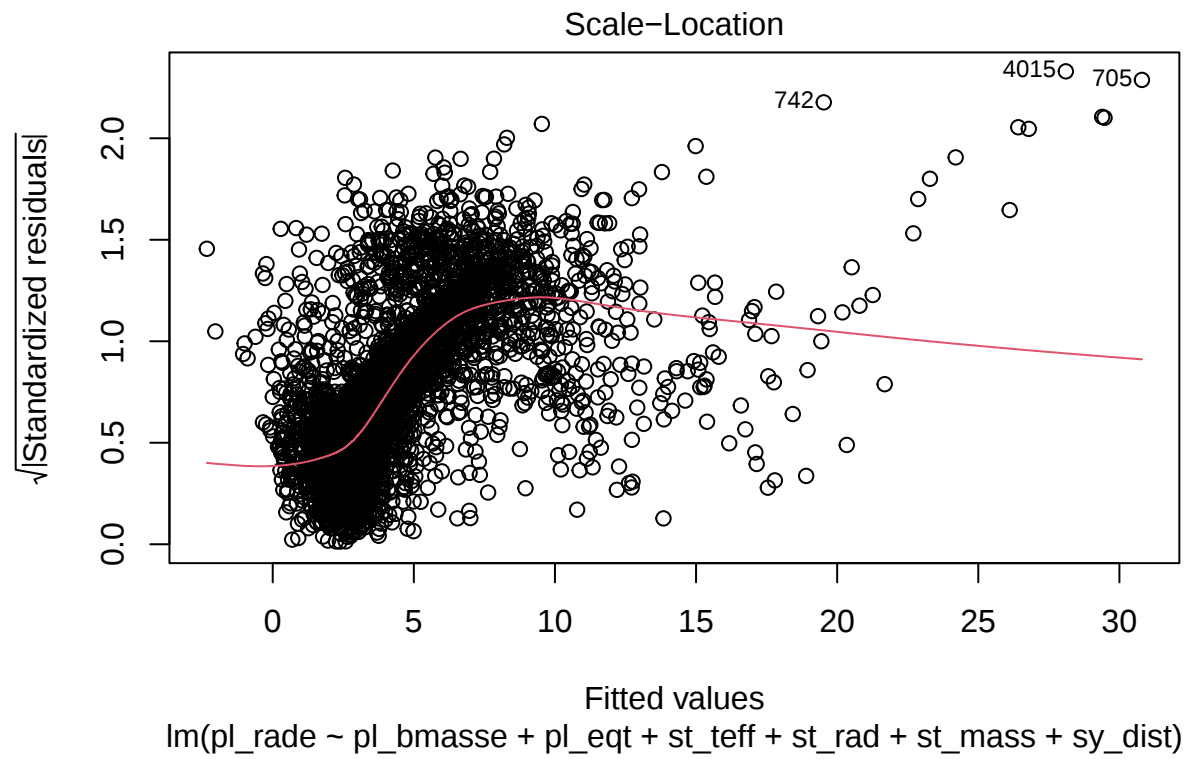
```

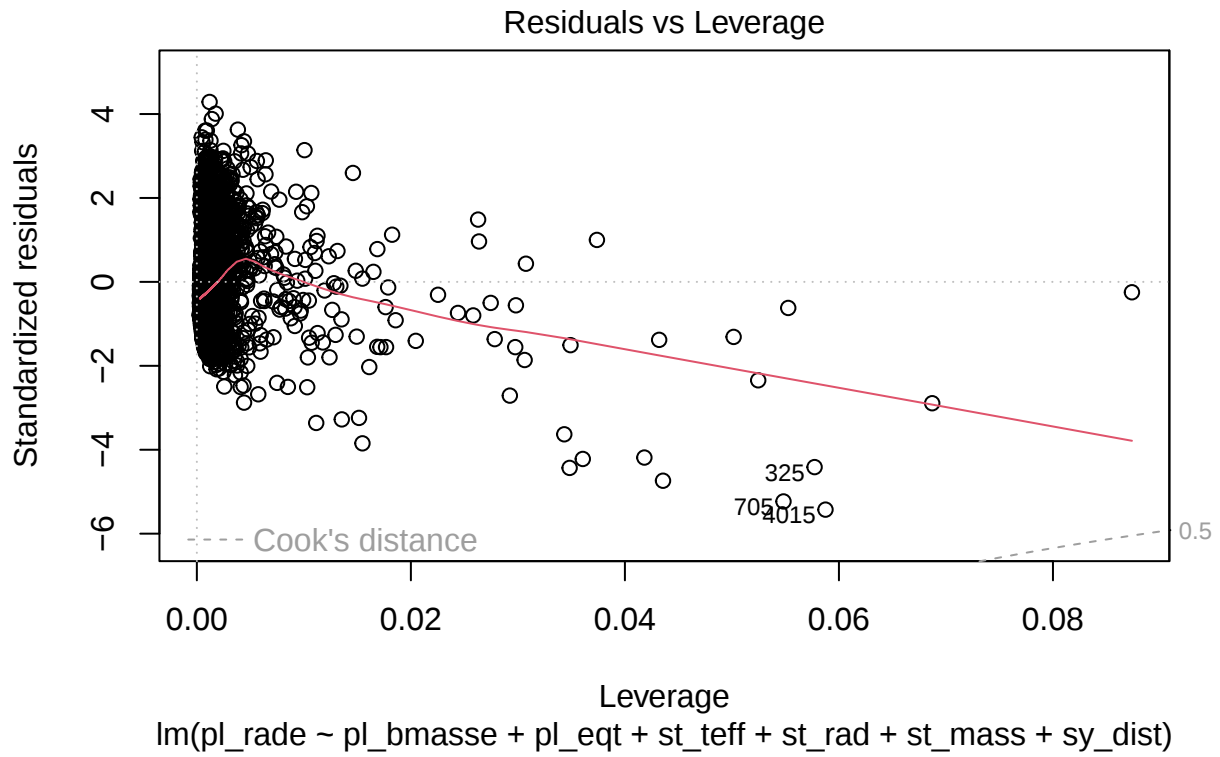




Theoretical Quantiles

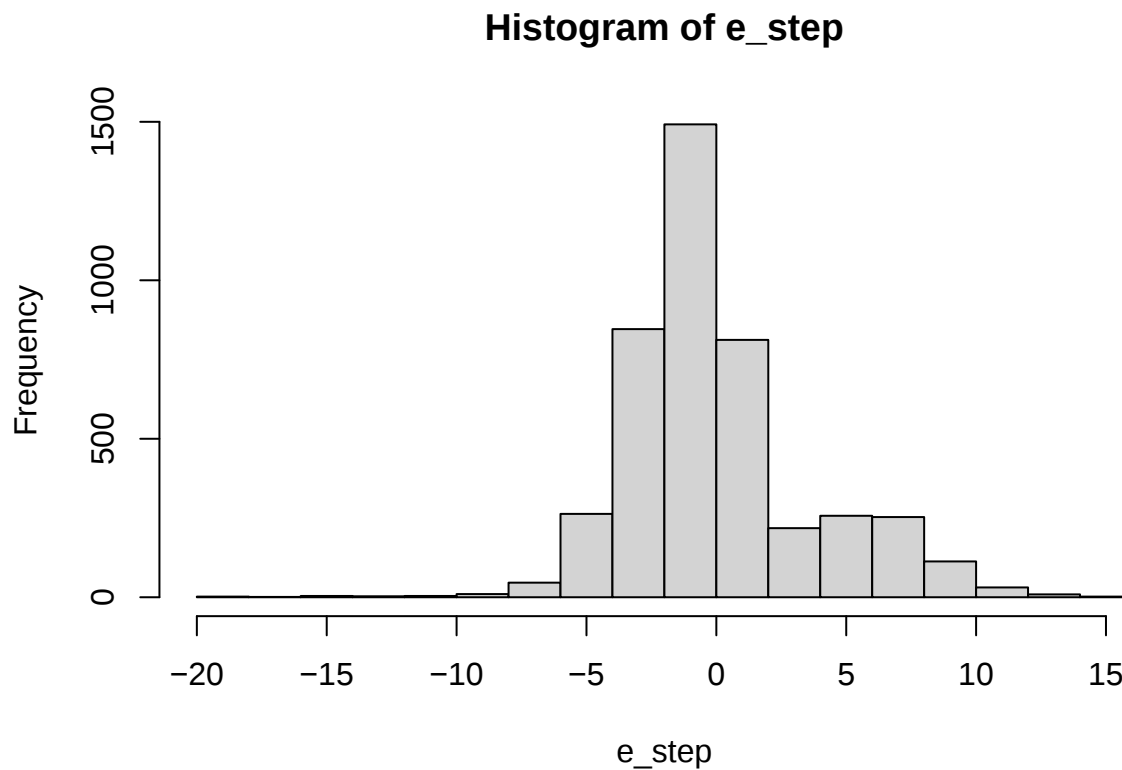
$\ln(\text{pl_rade}) \sim \text{pl_bmasse} + \text{pl_eqt} + \text{st_teff} + \text{st_rad} + \text{st_mass} + \text{sy_dist}$





```
# Residuos del modelo AIC
e_step <- residuals(fit_step)

# Histograma
hist(e_step)
```



```
# Normalidad  
shapiro.test(e_step)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: e_step  
## W = 0.93015, p-value < 2.2e-16
```

```
# Homocedasticidad  
bptest(fit_step)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: fit_step  
## BP = 1371.9, df = 6, p-value < 2.2e-16
```

```
# Autocorrelación  
dwtest(fit_step, alternative = "two.sided")
```

```
##  
## Durbin-Watson test  
##
```

```
## data: fit_step
## DW = 1.1061, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

```
test_residuos <- data.frame(
  Modelo = c("Modelo general", "Modelo AIC"),
  Shapiro_pvalor = c(
    shapiro.test(residuals(fit))$p.value,
    shapiro.test(residuals(fit_step))$p.value
  ),
  Breusch_Pagan_pvalor = c(
    bptest(fit)$p.value,
    bptest(fit_step)$p.value
  ),
  Durbin_Watson_pvalor = c(
    dwtest(fit, alternative = "two.sided")$p.value,
    dwtest(fit_step, alternative = "two.sided")$p.value
  )
)

test_residuos
```

```
##           Modelo Shapiro_pvalor Breusch_Pagan_pvalor Durbin_Watson_pvalor
## 1 Modelo general  3.988008e-41      1.874945e-291      9.399645e-192
## 2      Modelo AIC  3.836369e-41      2.998720e-293      3.819913e-192
```

Ambos modelos presentan métricas de bondad del ajuste prácticamente idénticas. El modelo AIC elimina `pl_orbsmax`, que no resultaba significativa, y obtiene un AIC ligeramente inferior (23604.94 frente a 23605.56), por lo que se prefiere el modelo simplificado por ser más parsimonioso sin perder capacidad explicativa.

Sin embargo, ninguno de los dos modelos permite realizar inferencia estadística de forma completamente válida, ya que ambos incumplen los supuestos básicos de normalidad, homocedasticidad e independencia de los residuos. Los p-valores y los intervalos de confianza deben interpretarse con cautela, ya que su validez depende del cumplimiento de dichos supuestos.

7 Conclusiones finales

En este trabajo se ha analizado un conjunto de datos de exoplanetas confirmados con el objetivo de estudiar qué variables físicas y orbitales se relacionan con el radio planetario (`pl_rade`). Para ello, se realizó primero un análisis exploratorio de los datos, posteriormente se ajustó un modelo de regresión lineal múltiple y, finalmente, se evaluó la calidad del ajuste y el cumplimiento de los supuestos del modelo.

El análisis exploratorio mostró que la variable dependiente presenta una distribución asimétrica y con valores extremos, algo esperable en una base de datos astronómica formada por planetas de características muy diferentes. Además, se detectó una relación muy elevada entre `pl_orbper` y `pl_orbsmax`, lo que indicaba un problema de multicolinealidad. Por este motivo, se decidió conservar `pl_orbsmax` y excluir `pl_orbper` del modelo inicial.

El modelo lineal múltiple resultó globalmente significativo, lo que indica que las variables explicativas consideradas aportan información relevante para explicar el radio planetario. En particular, variables como `pl_bmasse`, `pl_eqt`, `st_teff`, `st_rad`, `st_mass` y `sy_dist` resultaron estadísticamente significativas. La única variable que no mostró un efecto significativo en el modelo fue `pl_orbsmax`, lo que sugiere que, una vez controlado el efecto del resto de variables, el semieje mayor orbital no aporta evidencia suficiente para explicar linealmente el radio del planeta.

Desde el punto de vista interpretativo, los resultados indican que el radio planetario está especialmente relacionado con la masa del planeta y con características de la estrella anfitriona. El modelo con variables estandarizadas permitió comparar la importancia relativa de los predictores, destacando principalmente `st_mass`, `pl_bmasse` y `st_teff` como las variables con mayor impacto relativo sobre `pl_rade`.

No obstante, la capacidad explicativa del modelo es moderada. El valor de R^2 ajustado indica que el modelo explica aproximadamente el 39.5 % de la variabilidad del radio planetario, por lo que una parte importante de la variación queda sin explicar. Esto es razonable, ya que el tamaño de un exoplaneta depende de muchos factores físicos que no están incluidos en el modelo, como la composición interna, la densidad, la atmósfera, la edad del sistema o los procesos de formación y evolución planetaria.

Además, el análisis de residuos mostró que el modelo no cumple completamente los supuestos clásicos de regresión lineal. En particular, se detectaron problemas de normalidad, heterocedasticidad y autocorrelación en los residuos. Por tanto, aunque el modelo es útil como primera aproximación explicativa, sus resultados deben interpretarse con cautela y no deben considerarse como una herramienta predictiva definitiva.

La selección automática de variables mediante AIC permitió obtener un modelo más simple, manteniendo una capacidad explicativa similar. Sin embargo, los problemas detectados en los residuos indican que la mejora del modelo no debería centrarse únicamente en eliminar variables, sino también en considerar transformaciones o enfoques alternativos.

Como posibles mejoras futuras, sería recomendable probar transformaciones logarítmicas en algunas variables, especialmente en aquellas con distribuciones muy asimétricas, como `pl_rade`, `pl_bmasse`, `pl_orbsmax` o `sy_dist`. También podrían explorarse modelos no lineales, modelos robustos o técnicas de aprendizaje automático que capturen relaciones más complejas entre las características físicas de los exoplanetas y su radio.

En conclusión, el modelo permite identificar asociaciones relevantes entre el radio planetario y varias características del planeta y de su estrella anfitriona, especialmente la masa del planeta y la masa estelar. Sin embargo, debido a la complejidad física del fenómeno y a las limitaciones detectadas en los supuestos del modelo, los resultados deben entenderse como una aproximación estadística inicial y no como una explicación completa del tamaño de los exoplanetas.

8 Ampliación: modelo con transformación logarítmica

Como ampliación del análisis, se plantea un modelo alternativo basado en transformaciones logarítmicas. Esta propuesta se justifica por la elevada asimetría observada en algunas variables del conjunto de datos, especialmente en `pl_rade`, `pl_bmasse`, `pl_orbsmax` y `sy_dist`. Estas variables presentan valores muy dispersos y algunos valores extremos, lo que puede afectar al cumplimiento de los supuestos del modelo lineal clásico.

La transformación logarítmica puede ayudar a reducir la influencia de valores extremos, estabilizar parcialmente la varianza y mejorar la relación lineal entre la variable dependiente y algunas variables explicativas. Por este motivo, se ajusta un modelo donde se transforma logarítmicamente el radio planetario y algunas variables estrictamente positivas.

```
# Creamos una nueva base con transformaciones logarítmicas
datos_log <- datos %>%
  mutate(
    log_pl_rade = log(pl_rade),
    log_pl_bmasse = log(pl_bmasse),
    log_pl_orbsmax = log(pl_orbsmax),
    log_sy_dist = log(sy_dist)
  )

# Ajuste del modelo logarítmico
fit_log <- lm(
  log_pl_rade ~ log_pl_bmasse + log_pl_orbsmax + pl_eqt +
    st_teff + st_rad + st_mass + log_sy_dist,
  data = datos_log
)

# Resumen del modelo
summary(fit_log)
```

```
##
## Call:
## lm(formula = log_pl_rade ~ log_pl_bmasse + log_pl_orbsmax + pl_eqt +
##      st_teff + st_rad + st_mass + log_sy_dist, data = datos_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08251 -0.12595  0.00827  0.13445  1.52787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.116e-02  6.168e-02   0.505 0.613458
## log_pl_bmasse  3.772e-01  3.083e-03 122.354 < 2e-16 ***
## log_pl_orbsmax -5.114e-03  1.024e-02  -0.499 0.617532
## pl_eqt        -2.154e-05  2.475e-05  -0.870 0.384298
## st_teff       -2.454e-05  1.655e-05  -1.483 0.138181
## st_rad         7.574e-02  2.182e-02   3.470 0.000525 ***
## st_mass        2.322e-01  5.762e-02   4.029 5.69e-05 ***
## log_sy_dist   -5.867e-03  5.396e-03  -1.087 0.276976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.344 on 4358 degrees of freedom
## Multiple R-squared:  0.8347, Adjusted R-squared:  0.8345
## F-statistic: 3144 on 7 and 4358 DF,  p-value: < 2.2e-16
```

En este modelo, la variable dependiente es `log_pl_rade`, es decir, el logaritmo del radio planetario. Por tanto, la interpretación de los coeficientes cambia respecto al modelo lineal original.

En el caso de las variables también transformadas en logaritmos, como `log_pl_bmasse`, `log_pl_orbsmax` y `log_sy_dist`, los coeficientes pueden interpretarse aproximadamente como elasticidades. Es decir, indican el cambio porcentual esperado en el radio planetario ante un cambio porcentual en la variable explicativa, manteniendo constantes el resto de variables.

En cambio, para las variables que no se han transformado, como `pl_eqt`, `st_teff`, `st_rad` y `st_mass`, los coeficientes indican el cambio relativo aproximado en el radio planetario ante un aumento de una unidad en la variable explicativa.

8.1 Comparación de la bondad del ajuste

A continuación, se comparan los principales indicadores de bondad del ajuste del modelo general, el modelo seleccionado mediante AIC y el modelo con transformación logarítmica.

```
# Comparación de modelos mediante métricas principales

comparacion_modelos <- tibble(
  Modelo = c("Modelo general", "Modelo AIC", "Modelo logarítmico"),
  R2 = c(
    summary(fit)$r.squared,
    summary(fit_step)$r.squared,
    summary(fit_log)$r.squared
  ),
  R2_ajustado = c(
    summary(fit)$adj.r.squared,
    summary(fit_step)$adj.r.squared,
    summary(fit_log)$adj.r.squared
  ),
  AIC = c(
    AIC(fit),
    AIC(fit_step),
    AIC(fit_log)
  ),
  Sigma_residual = c(
    sigma(fit),
    sigma(fit_step),
    sigma(fit_log)
  )
)

knitr::kable(
  comparacion_modelos,
  digits = 4,
  caption = "Comparación de la bondad del ajuste entre modelos"
)
```

Table 2: Comparación de la bondad del ajuste entre modelos

Modelo	R2	R2_ajustado	AIC	Sigma_residual
Modelo general	0.3962	0.3952	23605.559	3.6084
Modelo AIC	0.3960	0.3951	23604.939	3.6085
Modelo logarítmico	0.8347	0.8345	3081.181	0.3440

El modelo logarítmico se compara con los modelos anteriores para comprobar si la transformación mejora la capacidad explicativa, reduce el AIC o disminuye la dispersión de los residuos. En particular, un mayor valor de R^2 ajustado, un menor AIC y una menor desviación típica residual indicarían una mejora relativa del ajuste.

No obstante, debe tenerse en cuenta que el modelo logarítmico no está expresado en la misma escala que el modelo original, ya que la variable dependiente pasa de ser `pl_rade` a `log_pl_rade`. Por ello, la comparación debe interpretarse con cautela, especialmente en el caso de la desviación típica residual.

8.2 Intervalos de confianza del modelo logarítmico

```
# Intervalos de confianza del modelo logarítmico
confint(fit_log)
```

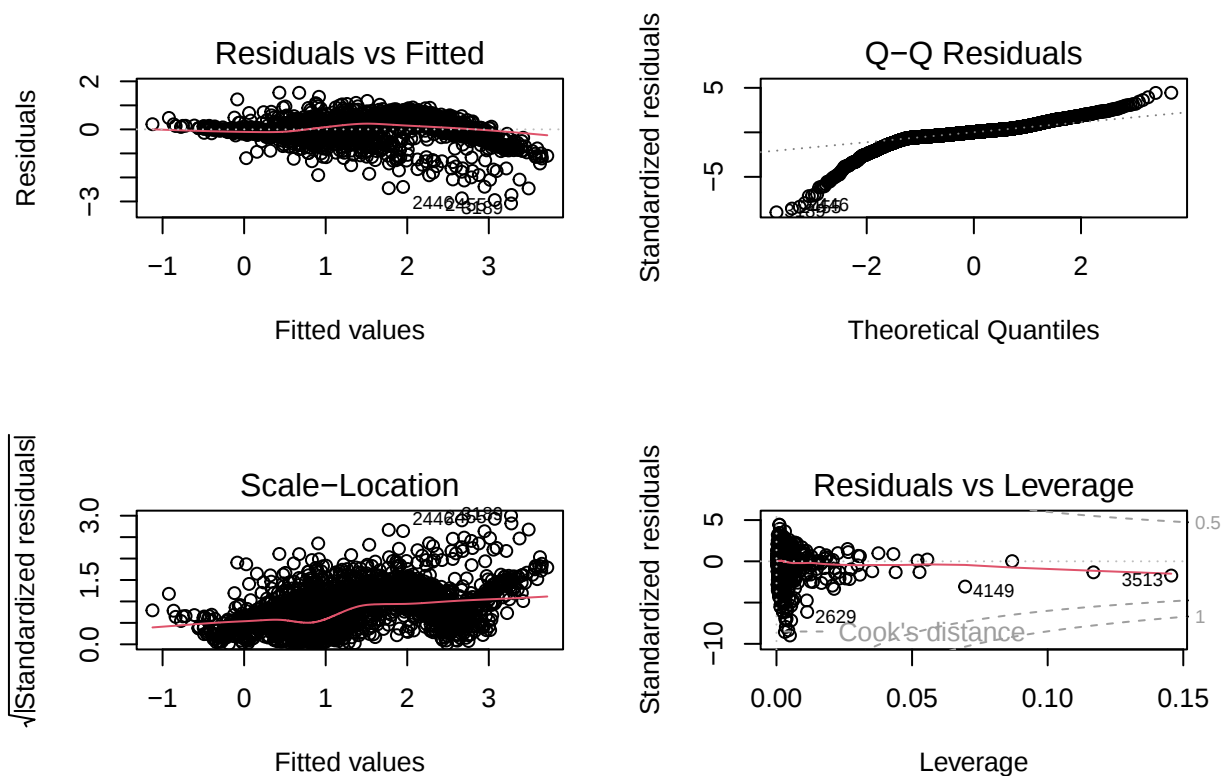
```
##                2.5 %      97.5 %
## (Intercept)  -8.975845e-02 1.520730e-01
## log_pl_bmasse 3.711205e-01 3.832072e-01
## log_pl_orbsmax -2.518869e-02 1.496151e-02
## pl_eqt       -7.007009e-05 2.699273e-05
## st_teff      -5.698024e-05 7.903804e-06
## st_rad       3.294918e-02 1.185225e-01
## st_mass      1.192021e-01 3.451204e-01
## log_sy_dist  -1.644491e-02 4.711703e-03
```

A partir de los intervalos de confianza al 95 %, se observa que `log_pl_bmasse`, `st_rad` y `st_mass` presentan efectos estadísticamente significativos, ya que sus intervalos no contienen el valor 0. En concreto, `log_pl_bmasse` muestra un intervalo claramente positivo, lo que confirma que la masa planetaria sigue siendo la variable más relevante para explicar el radio planetario en el modelo logarítmico. Por su parte, `st_rad` y `st_mass` también presentan intervalos positivos, lo que refuerza la importancia de las características de la estrella anfitriona.

En cambio, los intervalos de `log_pl_orbsmax`, `pl_eqt`, `st_teff` y `log_sy_dist` contienen el valor 0, por lo que no se puede afirmar que estas variables tengan un efecto estadísticamente significativo sobre el logaritmo del radio planetario al nivel de confianza del 95 %.

8.3 Diagnóstico de residuos del modelo logarítmico

```
# Gráficos diagnósticos del modelo logarítmico
par(mfrow = c(2, 2))
plot(fit_log)
```



```
par(mfrow = c(1, 1))
```

Los gráficos diagnósticos permiten comprobar visualmente si la transformación logarítmica mejora el comportamiento de los residuos. En concreto, se analiza si los residuos presentan una distribución más cercana a la normalidad, si la varianza es más constante y si se reduce la influencia de observaciones extremas.

```
# Contraste de normalidad de los residuos
shapiro.test(residuals(fit_log))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit_log)
## W = 0.8486, p-value < 2.2e-16
```

```
# Contraste de homocedasticidad
bptest(fit_log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit_log
## BP = 499.68, df = 7, p-value < 2.2e-16
```

```

# Contraste de autocorrelación
dwtest(fit_log)

##
## Durbin-Watson test
##
## data: fit_log
## DW = 1.4986, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

# Valores influyentes mediante distancia de Cook
cook_log <- cooks.distance(fit_log)
umbral_cook_log <- 4 / nrow(datos_log)

influyentes_log <- which(cook_log > umbral_cook_log)

length(influyentes_log)

## [1] 257

head(influyentes_log)

```

```

## 3 4 6 13 19 24
## 3 4 6 13 19 24

```

Para completar el diagnóstico, se aplican los mismos contrastes utilizados en los modelos anteriores: normalidad de los residuos, homocedasticidad, autocorrelación y detección de valores influyentes. De esta forma, se puede valorar si el modelo logarítmico corrige total o parcialmente las limitaciones detectadas en el modelo lineal inicial.

Los contrastes estadísticos muestran que el modelo logarítmico mejora el ajuste, pero sigue presentando problemas en los supuestos básicos. El test de Shapiro-Wilk rechaza la normalidad de los residuos, ya que el p-valor es inferior a 0.001. El test de Breusch-Pagan también rechaza la hipótesis de homocedasticidad, por lo que la varianza de los residuos no puede considerarse constante. Además, el test de Durbin-Watson detecta autocorrelación positiva en los residuos, con un estadístico inferior a 2 y un p-valor inferior a 0.001.

Por último, el análisis mediante distancia de Cook identifica 257 observaciones potencialmente influyentes. Esto confirma que, aunque la transformación logarítmica reduce parte de la asimetría y mejora notablemente la bondad del ajuste, todavía existen observaciones extremas o influyentes que afectan al comportamiento del modelo.

8.4 Conclusión de la ampliación

La transformación logarítmica mejora de forma clara la capacidad explicativa del modelo. Mientras que el modelo general y el modelo seleccionado mediante AIC explicaban aproximadamente el 39.5 % de la variabilidad del radio planetario, el modelo logarítmico alcanza un R^2 ajustado de 0.8345. Además, el AIC se reduce notablemente, pasando de valores cercanos a 23605 en los modelos anteriores a 3081.18 en el modelo logarítmico. Esto indica que, desde el punto de vista de la bondad del ajuste, la transformación logarítmica supone una mejora importante.

También se reduce considerablemente la desviación típica residual, que pasa de aproximadamente 3.61 en los modelos con la variable original a 0.344 en el modelo logarítmico. No obstante, esta comparación debe

interpretarse con cautela, ya que el modelo logarítmico no predice directamente el radio planetario en radios terrestres, sino su logaritmo.

En cuanto a los coeficientes del modelo logarítmico, las variables estadísticamente significativas son `log_pl_bmasse`, `st_rad` y `st_mass`. La variable con mayor importancia es `log_pl_bmasse`, cuyo coeficiente positivo indica que los planetas con mayor masa tienden a presentar un mayor radio, manteniendo constantes el resto de variables. Además, `st_rad` y `st_mass` también presentan efectos positivos y significativos, lo que sugiere que las características de la estrella anfitriona siguen siendo relevantes para explicar el tamaño planetario. En cambio, `log_pl_orbsmax`, `pl_eqt`, `st_teff` y `log_sy_dist` no resultan significativas al nivel del 5 %, ya que sus intervalos de confianza contienen el valor 0.

Sin embargo, el diagnóstico de residuos muestra que la transformación logarítmica no resuelve completamente las limitaciones del modelo. El test de Shapiro-Wilk presenta un p-valor inferior a 0.001, por lo que se rechaza la hipótesis de normalidad de los residuos. El test de Breusch-Pagan también presenta un p-valor inferior a 0.001, lo que indica que persisten problemas de heterocedasticidad. Además, el test de Durbin-Watson obtiene un estadístico de 1.4986 con p-valor inferior a 0.001, por lo que se detecta autocorrelación positiva en los residuos. Finalmente, mediante la distancia de Cook se identifican 257 observaciones potencialmente influyentes.

Por tanto, el modelo logarítmico puede considerarse una mejora importante respecto al modelo lineal original en términos de capacidad explicativa y ajuste global. No obstante, no cumple plenamente los supuestos clásicos de regresión lineal, por lo que sus resultados deben interpretarse con cautela. En conjunto, la transformación logarítmica permite capturar mejor la relación entre las variables astronómicas y el radio planetario, pero la complejidad y heterogeneidad de los datos sugieren que podrían ser necesarios modelos más flexibles, como regresión robusta, modelos no lineales o modelos específicos para poblaciones diferenciadas de exoplanetas.