



Universidad Austral
Facultad de Ingeniería
Maestría en Ciencia de Datos

Regresión Avanzada 2024
Trabajo Final
Detección de exoplanetas y sus características

Profesoras:

CHAN, Débora

OLIVA, Cecilia

Alumno:

NICOLAU, Jorge Enrique

2024

Resumen	3
Palabras clave.....	3
Abstract	3
Keywords	3
Detección de exoplanetas y sus características	4
El problema.....	4
Antecedentes	4
Trabajos relacionados	5
Análisis Exploratorio.....	5
El dataset de exoplanetas detectados con el telescopio espacial Kepler.....	6
Preprocesamiento.....	6
Revisión de variables numéricas.....	7
Detención de valores faltantes	9
Buscando correlaciones	10
Detección de valores atípicos (outliers) en las predictoras	11
Análisis de regresión.....	12
Regresión lineal univariada	12
Regresión lineal: radio del exoplaneta según pl_ratdor.....	12
Regresión lineal: radio del exoplaneta según pl_trandep	15
Regresión lineal: radio del exoplaneta según pl_rvamp	19
Regresión lineal multivariada.....	22
Regresión logística	23
Posibles mejoras para futuros análisis	26
Conclusiones.....	26
Referencias	28

Resumen

Este estudio emplea **regresión lineal y logística** para analizar datos de la misión **Kepler** y predecir tanto el **radio de los exoplanetas** como su **habitabilidad**. En la regresión lineal, se modela el radio planetario (**pl_radj**) en función de variables como la relación de radio estelar (**pl_ratdor**), la profundidad del tránsito (**pl_trandep**) y la amplitud de velocidad radial (**pl_rvamp**). Los resultados indican un ajuste significativo, explicando **el 84.25% de la variabilidad** en los datos, lo que sugiere que el modelo es una herramienta útil para estimar el tamaño de los exoplanetas.

Por otro lado, la regresión logística se utiliza para definir un **Índice de Habitabilidad Planetaria (PHI)** basado en factores como temperatura de equilibrio (**pl_eqt**), insolación (**pl_insol**) y semieje mayor (**pl_orbsmax**). Sin embargo, la falta de datos completos reduce la muestra a **solo 17 observaciones**, lo que genera **sobreajuste y alta incertidumbre** en las predicciones. Para mejorar la calidad del modelo, se recomienda aplicar **técnicas de imputación de datos** y explorar enfoques de **aprendizaje automático**, lo que permitiría mejorar la estabilidad y confiabilidad de los resultados.

Palabras clave

Exoplanetas, regresión lineal, regresión logística, habitabilidad planetaria, Índice de Habitabilidad Planetaria (PHI), modelo predictivo, Kepler/K2, aprendizaje automático, valores faltantes, sobreajuste.

Abstract

This study applies **linear and logistic regression** to analyze data from the **Kepler mission** and predict both the **radius of exoplanets** and their **habitability**. In the linear regression model, planetary radius (**pl_radj**) is estimated based on variables such as stellar radius ratio (**pl_ratdor**), transit depth (**pl_trandep**), and radial velocity amplitude (**pl_rvamp**). The results indicate a strong fit, explaining **84.25% of the variability**, suggesting that the model is a valuable tool for estimating exoplanet size.

On the other hand, logistic regression is used to define a **Planetary Habitability Index (PHI)** based on equilibrium temperature (**pl_eqt**), insolation (**pl_insol**), and semi-major orbital axis (**pl_orbsmax**). However, the lack of complete data reduces the sample to **only 17 observations**, leading to **overfitting and high uncertainty** in the predictions. To improve model performance, it is recommended to apply **data imputation techniques** and explore **machine learning approaches**, which could enhance the stability and reliability of the results.

Keywords

Exoplanets, linear regression, logistic regression, planetary habitability, Planetary Habitability Index (PHI), predictive model, Kepler/K2, machine learning, missing values, overfitting.

Detección de exoplanetas y sus características

El problema

Durante casi tres décadas de investigación, la NASA ha confirmado la existencia de más de 5,600 exoplanetas en 4,151 sistemas planetarios. Estos mundos, conocidos como exoplanetas o planetas extrasolares, orbitan estrellas distintas al Sol. Aunque los astrónomos los imaginaron durante milenios, no fue hasta mediados de la década de 1990 cuando comenzaron a aparecer en los registros científicos.

Aproximadamente dos tercios de los exoplanetas descubiertos hasta ahora provienen del telescopio espacial Kepler, mientras que cientos más han sido identificados gracias a la misión TESS y otros observatorios terrestres. Estas misiones han generado una enorme cantidad de datos, con cerca de 10,000 posibles exoplanetas aún pendientes de confirmación. Un número impresionante si consideramos que hasta principios de los años noventa no se conocía ninguno, pero diminuto en comparación con los cientos de miles de millones que podrían existir solo en la Vía Láctea.

Para detectar estos mundos, explica Nardi (Nardi, 2024), los astrónomos emplean diversas técnicas, muchas de ellas llevadas al límite de la tecnología disponible en los observatorios espaciales: **astrometría** o detección a través del movimiento estelar como refiere Dong-Hong (Dong-Hong, 2023), **velocidades radiales** o detección a través del efecto Doppler como se expone en el artículo de Marín (Marín, 2018), **método del tránsito** o detección a través de la sombra planetaria según lo expuesto en la tesis de Alonso (Alonso Sobrino, 2006), **microlente gravitacional** o detección a través de la curvatura del espacio-tiempo como expone también Marín (Marín, 2011), **observación directa** o fotografía en medio del brillo estelar según lo reportado por NASA (NASA, 2022) y detección por cómo afectan la **rotación de púlsares** como por ejemplo en el trabajo pionero de Wolszczan (Wolszczan & Frail, 1992).

El análisis de exoplanetas presenta varios desafíos en la modelización de datos. La **multicolinealidad** es un problema común, ya que muchas variables orbitales, estelares y planetarias están altamente correlacionadas debido a su **origen físico común**, como la relación entre el **período orbital y el semieje mayor**, descrita por la **Ley de Kepler**. Además, el **ruido y la variabilidad** en los datos pueden afectar la precisión de las predicciones, ya que las mediciones suelen presentar **incertidumbre**, especialmente en la estimación de **masas y radios planetarios**, lo que refleja la diversidad natural de los sistemas estelares. Otro aspecto clave es la **no linealidad** en las relaciones entre propiedades estelares, orbitales y planetarias, como la **dependencia cúbica** entre el **semieje mayor y el período orbital** según la Ley de Kepler. Finalmente, la **alta dimensionalidad** de los datos, generada por la gran cantidad de sensores y metodologías para detectar exoplanetas, puede introducir **variables irrelevantes o redundantes**, aumentando innecesariamente la **complejidad del modelo**, por lo que es fundamental aplicar técnicas de selección de variables para mejorar la eficiencia del análisis.

Antecedentes

La detección de exoplanetas es un campo de investigación en constante evolución, con nuevos descubrimientos y técnicas emergentes que amplían nuestro conocimiento del universo. La ciencia de datos juega un papel fundamental en este proceso, permitiendo a los astrónomos analizar grandes volúmenes de datos y extraer información relevante sobre los exoplanetas y sus características.

Incluso **NASA** anima a la detección amateur de exoplanetas a través del sitio Exoplanet Watch, Brachman (Brachman, 2024) ofrece una guía detallada sobre cómo analizar observaciones de tránsitos de exoplanetas para generar curvas de luz, que representan las variaciones en el brillo de una estrella cuando un planeta pasa frente a ella.

Trabajos relacionados

Hay varios antecedentes del uso de aprendizaje automático y métodos estadísticos para la detección de exoplanetas

En el trabajo de Malik (Malik & Obermeier, 2021) se presenta una nueva técnica basada en **machine learning** para detectar exoplanetas mediante el **método de tránsito**. Se empleó la biblioteca **TSFresh** para extraer **789 características** de curvas de luz y entrenar un clasificador de **gradient boosting** con **LightGBM**.

En el paper de Cambazard y otros (Cambazard, Catusse, Chomez, & Lagrange, 2025) se trata la problemática de **detección de imagen directa de exoplanetas** y cómo separar el ruido de fondo de las señales planetarias. Los métodos estadísticos recientes evitan la auto sustracción de señales de interés, a diferencia del enfoque inicial basado en **imágenes diferenciales angulares (ADI)**. Sin embargo, estos métodos pueden generar **muchos falsos positivos** si no se establecen umbrales conservadores, lo que a su vez puede hacer que se pierdan exoplanetas débiles. Este estudio extiende un marco estadístico incorporando una **regresión logística** para filtrar candidatos, utilizando características ópticas en dos longitudes de onda. Se aplica **detección de bordes y algoritmos de clustering** para procesar sub-imágenes.

Cardenas y otros (Cardenas, Lozano, Marquez, Torres, & Delgado-Correal, 2022) desarrolla un ensamblado de **software autónomo** para detectar tránsitos planetarios en **curvas de luz estelares**, utilizando un **clasificador de lógica difusa** y evitando la búsqueda manual de tránsitos en grandes volúmenes de datos.

En el trabajo de Venkata y otros (Venkata, Jahnavi, Venkata, & Muvva, 2023) se presenta una técnica para detectar exoplanetas **usando el método de tránsito**. El objetivo es mejorar las técnicas tradicionales en astronomía con el uso de algoritmos de aprendizaje automático. Para ello, utilizan seis modelos diferentes de aprendizaje automático: **Random Forest, Decision Tree, Support Vector Classifier, K-Nearest Neighbor y Multi-Layer Perceptron**. Al comparar las precisiones concluye que combinando cuatro de esos modelos (Support Vector Classifier, K-Nearest Neighbor, Random Forest y Multi-Layer Perceptron) usando **bagging**, se lograba una mayor precisión.

Pimentel y otros (Pimentel, Amorim, & Rudzicz, 2024) se centran en detectar posibles planetas utilizando el método de tránsito con un enfoque novedoso al clasificar estrellas mediante un conjunto de características extraídas de series temporales en tres dominios: **temporal, estadístico y espectral**. Estas características se utilizan para entrenar y evaluar modelos con datos del **telescopio Kepler**, y los resultados superan algunos enfoques existentes. Además, el proceso de **validación cruzada** se emplea para eliminar sesgos y evaluar mejor los modelos.

Inspirándose en estos antecedentes, este trabajo busca **utilizar la regresión parámetros faltantes de los exoplanetas** así como **predecir su habitabilidad**.

Análisis Exploratorio

En el reconocido libro de Tuckey (Tukey, 1977) se explica que el análisis exploratorio de datos (EDA) se enfoca en descubrir patrones, anomalías y relaciones dentro de los datos, utilizando métodos visuales y estadísticos. En su base, implica abordar la incertidumbre y la diversidad de estructuras posibles en los datos, evitando asumir un único modelo probabilístico como absoluto. El EDA enfatiza la exploración sobre la inferencia estricta, buscando métodos robustos y flexibles que funcionen en diversas circunstancias, con el objetivo de proporcionar una comprensión preliminar y preparar el terreno para análisis más profundos.

El dataset de exoplanetas detectados con el telescopio espacial Kepler

El Archivo de Exoplanetas de la NASA es un catálogo astronómico en línea y un servicio de datos que recopila, correlaciona y organiza información sobre exoplanetas y sus estrellas anfitrionas. Además, proporciona herramientas para trabajar con estos datos. IPAC (IPAC, 2021) ofrece un archivo que está dedicado a la recopilación y difusión de conjuntos de datos públicos claves utilizados en la búsqueda y caracterización de planetas extrasolares y sus estrellas.

Al momento de la extracción (14 de enero de 2025) el dataset de exoplanetas contiene información sobre 28.217 exoplanetas (entre confirmados, propuestos y en estudio) y sus características. La base de datos de IPAC contiene información sobre los exoplanetas detectados por la misión Kepler y otros telescopios, pero para este trabajo se utilizará solo la información de los exoplanetas detectados por la misión Kepler. Para más detalle sobre la misión referirse al artículo Web de Cermak (Cermak, 2024).

Preprocesamiento

Se cargarán los datos, se filtrarán las filas de los exoplanetas confirmados, las columnas relevantes, se eliminarán los valores faltantes y se realizará un análisis exploratorio de los datos para luego realizar la regresión lineal y logística. Los datos iniciales están en **keplerexoplanets.csv** en el repositorio creado por Nicolau (Nicolau, 2025) para este trabajo. El mismo se obtuvo de la página de IPAC (IPAC, 2025) dependiente del Caltech.

El dataset contiene información sobre los sistemas exoplanetarios detectados por la misión Kepler y el diccionario de datos se encuentra en línea en IPAC (IPAC, 2024) donde se explica en más profundidad la semántica de las variables.

Para el análisis se considerarán solo los **exoplanetas confirmados y se eliminarán las columnas no relevantes para el análisis**.

Para reducir la dimensionalidad del problema, a continuación se **eliminan las columnas con referencias a sitios web no relevantes para el análisis**. También se eliminan los **datos de referencia de los planetas, estrellas y sistemas no relevantes** para el análisis tales como referencias a catálogos externos de estrellas y sistemas, **información de publicación del descubrimiento, información de detección, información de fotometría**, banderas de organización interna del dataset, información del sistema planetario y detalles técnicos de la detección.

Para reducir la cantidad de registros del dataset y como se busca predecir la habitabilidad de los exoplanetas en sistemas similares al Solar, **se filtran los sistemas con estrellas de tamaño similar al Sol (0.9 a 1.1 radios solares)**. Además se considerarán para la muestra **solo los exoplanetas publicados en la literatura científica**, por lo que se filtra por “Published Confirmed” en la columna **soltype**.

El dataset resultante del preprocesado **contiene 2547 registros y 68 columnas** y tiene estas columnas:

- **Posición y Coordenadas.** La ubicación de la estrella en el cielo se expresa en diferentes sistemas de coordenadas: la **ascensión recta (ra, rastr)** y la **declinación (dec, decstr)** en coordenadas ecuatoriales, además de las **coordenadas galácticas (glat, glon)** y **eclípticas (elat, elon)**. También se incluye la **distancia al sistema (sy_dist)**, su **paralaje (sy_plx)** y su **movimiento propio (sy_pm, sy_pmra, sy_pmdec)**.
- **Identificación del Sistema Estelar.** El sistema estelar se puede identificar mediante diversos nombres provenientes de catálogos astronómicos, como el **Henry Draper (hd_name)**, **Hipparcos (hip_name)** y el **nombre comúnmente usado en la literatura científica (hostname)**.

- **Propiedades del Planeta.** Los exoplanetas tienen diversas características físicas y orbitales. Su **masa** se proporciona en diferentes unidades, como masas terrestres (**pl_masse**, **pl_bmasse**, **pl_cmasse**) y jovianas (**pl_massj**, **pl_bmassj**, **pl_cmassj**). Se incluyen valores de masa obtenidos por diferentes métodos, como la **mínima proyectada** (**pl_cmasse**, **pl_cmassj**) y la determinada por **velocidad radial** (**pl_msinie**, **pl_msinij**). También se reportan la **densidad** (**pl_dens**), el **radio** (**pl_rade**, **pl_radj**), y su relación con el radio estelar (**pl_ratdor**, **pl_rator**). El comportamiento orbital del planeta se describe con la **excentricidad** (**pl_orbeccen**), la **inclinación orbital** (**pl_orbincl**), el **argumento del periastro** (**pl_orblper**), el **semieje mayor** (**pl_orbsmax**), el **periodo orbital** (**pl_orbper**), y el **momento del periastro** (**pl_orbtper**). También se incluye información sobre su **insolación** (**pl_insol**), la **temperatura de equilibrio** (**pl_eqt**), y parámetros relacionados con su tránsito, como la **profundidad del tránsito** (**pl_trandep**), su **duración** (**pl_trandur**) y la **amplitud de velocidad radial** (**pl_rvamp**).
- **Propiedades de la Estrella.** Las estrellas que albergan exoplanetas están descritas en términos de su **edad** (**st_age**), **densidad** (**st_dens**), **gravedad superficial** (**st_logg**), **luminosidad** (**st_lum**), **masa** (**st_mass**), y **radio** (**st_rad**). También se incluyen medidas de su **metalicidad** (**st_met**, **st_metratio**), su **temperatura efectiva** (**st_teff**), y su **tipo espectral** (**st_spectype**). La información sobre su **movimiento y rotación** se expresa mediante la **velocidad radial** (**st_radv**), la **velocidad de rotación** (**st_vsin**), y el **periodo de rotación** (**st_rotp**).
- **Información del Sistema Planetario.** El número de cuerpos en el sistema se indica con **sy_pnum** para planetas y **sy_snum** para estrellas. También se incluye el **método de descubrimiento** (**discoverymethod**) y un **flag de tránsito de tiempo variante** (**ttv_flag**). En sistemas binarios, el parámetro **cb_flag** señala si el planeta orbita un sistema estelar doble.

Revisión de variables numéricas

Variable	Media	Des.Est.	Mín.	P25	Mediana	P75	Máx.	Hist.
cb_flag	0.0024	0.0485	0.0000	0.0000	0.0000	0.0000	1.0000	
dec	44.2366	3.6695	36.5773	41.3738	44.3155	47.1127	52.1491	
elat	64.7746	3.6360	57.6374	61.8480	64.7690	67.5246	72.4913	
elon	307.9950	8.0410	288.7515	302.3200	308.2267	314.3187	323.9048	
glat	12.9778	3.4193	5.9607	10.4051	12.6586	15.7959	21.1430	
glon	76.3718	3.6938	68.2330	73.3769	76.4011	79.3326	84.3822	
pl_bmasse	328.4638	1965.7298	0.7600	6.0133	18.4000	95.1000	25426.4000	
pl_bmassj	1.0335	6.1849	0.0024	0.0190	0.0580	0.2992	80.0000	
pl_dens	4.2419	8.2500	0.0300	0.6950	2.1300	5.5250	77.7000	
pl_eqt	938.0333	564.4973	251.0000	438.7500	861.5000	1146.5000	2188.0000	
pl_imppar	0.3940	0.2883	0.0000	0.1500	0.3300	0.6200	1.4830	
pl_insol	299.3455	598.0846	0.5900	26.9900	96.9450	304.9000	4849.2600	
pl_masse	328.4638	1965.7298	0.7600	6.0133	18.4000	95.1000	25426.4000	
pl_massj	1.0335	6.1849	0.0024	0.0190	0.0580	0.2992	80.0000	
pl_msinie	47.8000	NA	47.8000	47.8000	47.8000	47.8000	47.8000	
pl_msinij	0.1500	NA	0.1500	0.1500	0.1500	0.1500	0.1500	
pl_orbeccen	0.1177	0.1778	0.0000	0.0105	0.0420	0.1350	0.8380	
pl_orbincl	88.6474	1.7458	82.2140	87.8060	89.1855	89.7905	93.1500	
pl_orblper	146.9116	142.8866	-163.0000	49.2755	154.7000	261.6000	357.0300	
pl_orbper	31.4997	62.6948	0.5383	5.6992	13.0314	33.6013	1322.3000	
pl_orbsmax	0.1834	0.2236	0.0168	0.0655	0.1091	0.2170	2.4200	
pl_orbtper	2455003.75	106.5838	2454935.80	2454950.26	2454958.21	2455011.71	2455162.80	
pl_projbliq	-35.1429	60.4109	-135.0000	-61.5000	0.0000	4.0000	4.0000	
pl_rade	2.9323	2.4308	0.4000	1.6230	2.3700	3.0240	30.8000	
pl_radj	0.2616	0.2169	0.0360	0.1450	0.2110	0.2700	2.7480	

Variable	Media	Des.Est.	Mín.	P25	Mediana	P75	Máx.	Hist.
pl_ratdor	60.9640	83.9114	3.1000	11.5600	30.3000	78.3000	576.7000	
pl_rator	0.0288	0.0263	0.0056	0.0147	0.0213	0.0286	0.2873	
pl_rvamp	47.4204	88.2590	0.2800	2.1000	3.7200	71.9000	419.5000	
pl_trandep	0.1813	0.3624	0.0060	0.0481	0.0705	0.1063	2.2620	
pl_trandur	4.3811	2.2576	0.8244	2.8100	3.8212	5.4291	18.8860	
pl_tranmid	2455003.75	121.7668	2454832.90	2454967.11	2454973.96	2455004.68	2457959.96	
ra	291.3112	4.6829	280.2066	287.6980	291.4314	295.0027	301.5430	
st_age	4.5799	1.7720	0.1050	3.8000	4.2700	4.7900	11.9000	
st_dens	1.6002	1.3618	0.0044	1.2072	1.4497	1.6696	19.9316	
st_logg	4.4442	0.0556	4.2100	4.4100	4.4500	4.4800	5.0000	
st_lum	-0.0616	0.1126	-0.4090	-0.1360	-0.0630	0.0280	0.2200	
st_mass	0.9901	0.0642	0.6900	0.9500	0.9900	1.0300	1.2500	
st_met	0.0199	0.1570	-0.8160	-0.0400	0.0200	0.1000	0.4800	
st_rad	0.9922	0.0619	0.9000	0.9400	0.9900	1.0500	1.1000	
st_radv	-34.3600	29.8608	-98.9300	-57.1600	-24.7600	-20.9300	9.9600	
st_rotp	12.2225	5.6116	4.6900	10.1775	11.8900	14.3225	22.0500	
st_teff	5684.9878	236.0809	4388.3900	5549.5000	5688.0000	5833.5000	6484.0000	
st_vsin	2.2671	1.9863	0.3000	0.5000	2.0000	3.0000	10.4000	
sy_dist	875.6348	389.0261	68.1730	625.3390	860.3910	1101.1000	2879.8300	
sy_plx	1.4880	1.2326	0.3269	0.8774	1.1298	1.5694	14.6396	
sy_pm	10.9292	9.0598	0.1586	4.5898	8.4847	14.5826	77.6183	
sy_pmdec	-3.2419	11.7992	-66.6869	-9.1002	-3.1269	2.5670	48.3025	
sy_pmra	-0.3364	7.1920	-49.3144	-3.5431	-0.4287	3.0154	32.8344	
sy_pnum	2.0628	1.2462	1.0000	1.0000	2.0000	3.0000	6.0000	
sy_snum	1.0271	0.1624	1.0000	1.0000	1.0000	1.0000	2.0000	
ttv_flag	0.1496	0.3567	0.0000	0.0000	0.0000	0.0000	1.0000	

En cuanto a las coordenadas celestes, las distribuciones de **dec** y **ra** son relativamente centradas, mientras que **elon** presenta un rango amplio (288°-324°), abarcando las constelaciones de Capricornio, Acuario y Piscis. La latitud eclíptica **elat** es alta (57°-72°), lo que indica que los objetos están lejos del plano de la eclíptica y se concentran en la región del Polo Norte Eclíptico, cerca de la constelación del Dragón.

Las variables de masa planetaria (**pl_bmasse**, **pl_masse**, **pl_bmassj**, **pl_massj**) muestran distribuciones sesgadas a la derecha con valores extremos, lo que indica la presencia de algunos planetas muy masivos. De manera similar, las distribuciones de radio (**pl_rade**, **pl_radj**) están sesgadas, con la mayoría de los planetas siendo más grandes que la Tierra (mediana de **pl_rade** \approx 2.37). La excentricidad orbital (**pl_orbecen**) es predominantemente baja, sugiriendo órbitas casi circulares, mientras que la inclinación orbital (**pl_orbincl**) está centrada en 90°, lo que es consistente con la detección por el método de tránsito. El período orbital (**pl_orbper**) tiene una gran variabilidad, con una mediana de 13 días, pero valores extremos de hasta 1,322 días. La temperatura de equilibrio planetaria (**pl_eqt**) muestra una amplia distribución (251 K - 2,188 K), lo que indica la presencia de tanto planetas fríos como extremadamente calientes. La densidad (**pl_dens**) es variable, con una mediana de 2.13 g/cm³, reflejando la diversidad en la composición planetaria.

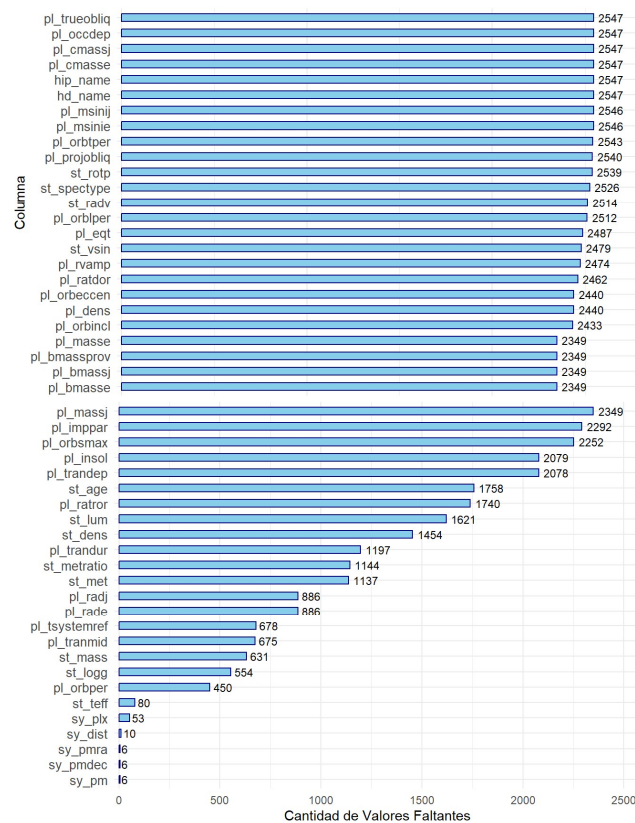
Respecto a las estrellas, la mayoría tienen masas y radios similares al Sol (**st_mass**, **st_rad** \approx 1.0), con temperaturas efectivas (**st_teff**) cercanas a 5,685 K, típicas de estrellas tipo G. La gravedad superficial (**st_logg**) está centrada en 4.45, lo que es característico de estrellas en la secuencia principal. La luminosidad (**st_lum**) también es similar a la del Sol en su mayoría.

El análisis de distancias muestra que los sistemas están relativamente lejos, con una mediana de 860 parsecs ($\approx 2,804$ años luz). El movimiento propio (**sy_pm**, **sy_pmra**, **sy_pmdec**) está mayormente centrado en 0, aunque algunas estrellas tienen movimientos más rápidos. El número de planetas por sistema (**sy_pnum**) tiene una mediana de 2, con un máximo de 6 planetas detectados.

Finalmente, los parámetros de tránsito (**pl_trandep**, **pl_trandur**) muestran distribuciones sesgadas, indicando que la mayoría de los tránsitos son poco profundos y de corta duración. La oblicuidad (**pl_projbliq**) es mayormente 0°, pero hay valores extremos de hasta -135°. En general, las distribuciones presentan sesgos y valores extremos, reflejando tanto la diversidad de los exoplanetas como la presencia de valores atípicos. La mayoría de los exoplanetas en el conjunto de datos son relativamente pequeños, aunque existen algunos extremadamente grandes, y las estrellas tienden a ser similares al Sol con variaciones en edad y metalicidad, en parte debido al filtro aplicado en el preprocesamiento.

Detención de valores faltantes

Se analiza la cantidad de valores faltantes en el dataset para identificar posibles problemas de calidad de datos.

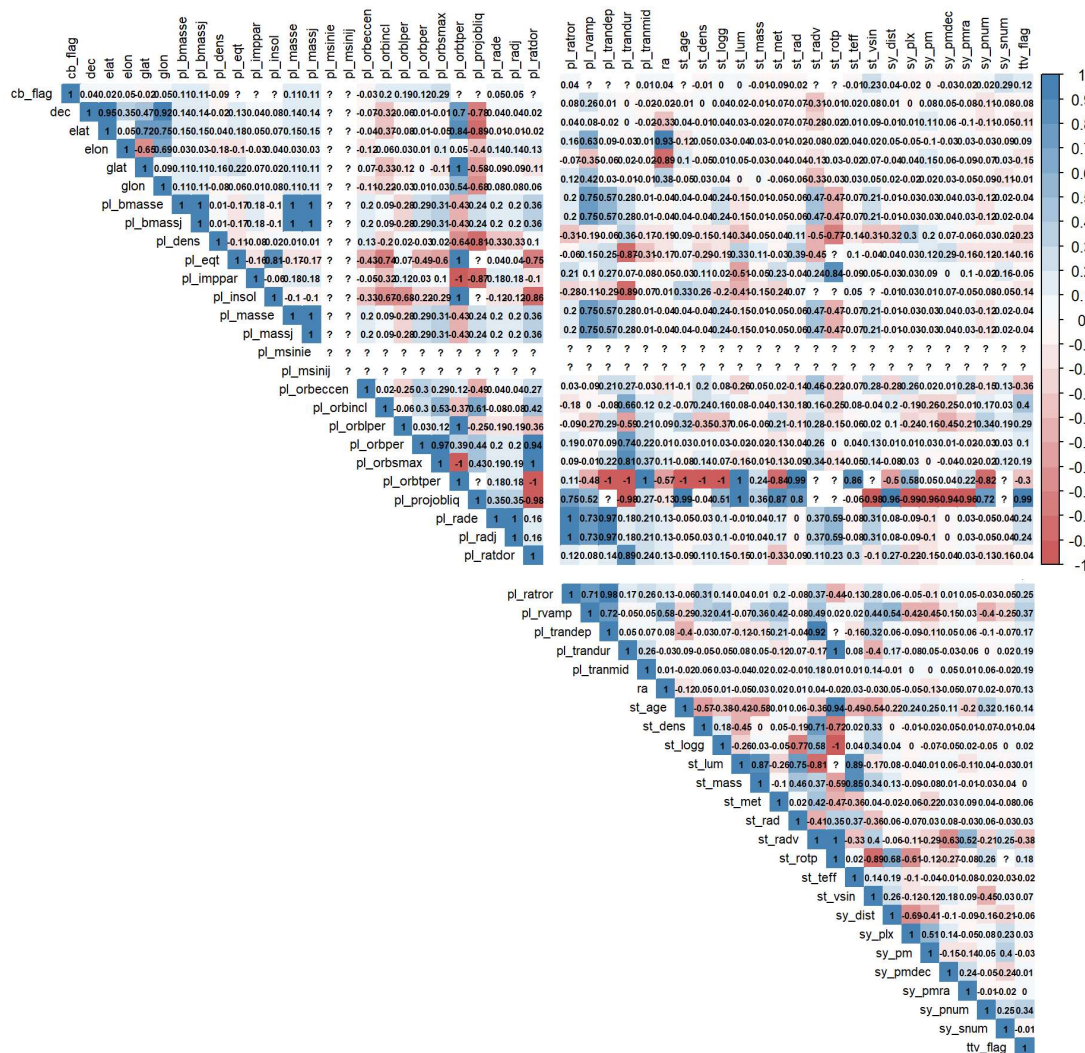


Como parte del proceso, se analizarán las **variables conocidas de los exoplanetas** para identificar aquellas que puedan actuar como **predictoras** de las variables con valores faltantes. Por ejemplo, en el caso del **radio del exoplaneta** (representado por **pl_radj** en radios jovianos y **pl_rade** en radios terrestres), que presenta **886 valores faltantes**, es posible estimarlo utilizando otras variables disponibles en el dataset. Algunas de estas incluyen el **período orbital en días** (**pl_orbper**, con **450 valores faltantes**), el **tiempo de conjunción** (**pl_tranmid**, con **675 valores faltantes**) y el

logaritmo en base 10 de la gravedad superficial de la estrella del sistema planetario (st_logg, con 554 valores faltantes).

Buscando correlaciones

Se analiza la correlación entre las variables numéricas del dataset para identificar posibles relaciones entre ellas utilizando los métodos de Pearson (detección de correlaciones lineales). A través de esta matriz, se pueden detectar tanto **correlaciones positivas** (cuando un aumento en una variable está asociado con un aumento en otra) como **correlaciones negativas** (cuando un aumento en una variable está relacionado con una disminución en otra). Estas relaciones pueden ser útiles para entender patrones en los datos y seleccionar las mejores variables predictoras para modelos de estimación.



Un ejemplo de estas correlaciones se observa en la relación entre el **cociente del semieje orbital mayor y el período orbital (pl_ratdor)**, el cual muestra una **correlación positiva** con el **período orbital (pl_orbper)**. Esto sugiere que a medida que el período orbital aumenta, también lo hace este cociente. En contraste, **pl_ratdor** presenta una **correlación negativa** con la **temperatura de equilibrio del exoplaneta (pl_eqt)**, lo que indica que los exoplanetas con períodos orbitales largos

tienden a tener temperaturas más bajas, lo cual es esperable, ya que suelen estar más alejados de su estrella anfitriona.

En cuanto a la variable dependiente que se seleccionó como **variable objetivo** en el análisis previo, el **radio del exoplaneta (pl_radj)**, se observan relaciones significativas con otras variables. Específicamente, tiene una **correlación positiva** con la **profundidad del tránsito (pl_trandep)**, lo que indica que los exoplanetas más grandes generan una mayor disminución en el brillo de su estrella al pasar frente a ella. Por otro lado, **pl_radj** muestra una **correlación negativa** con la **gravedad superficial de la estrella anfitriona (st_logg)**, lo que sugiere que los exoplanetas más grandes tienden a encontrarse en sistemas con estrellas de menor gravedad superficial, posiblemente porque estos sistemas albergan planetas de mayor tamaño.

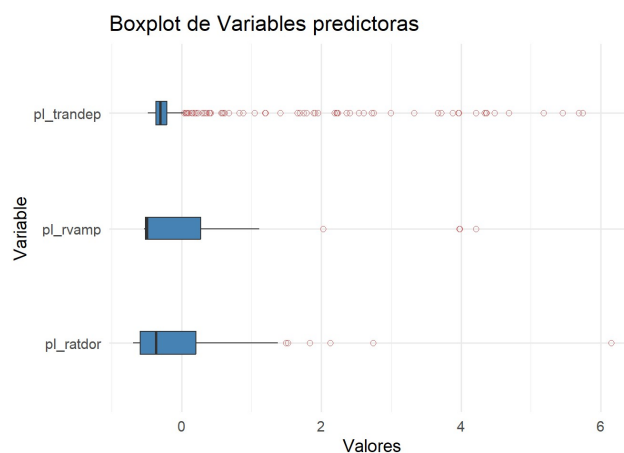
Más allá de la relación lineal evidente entre las medidas del **radio del exoplaneta en diferentes unidades** —es decir, **radios jovianos** y **radios terrestres**—, el análisis de correlación revela la presencia de otras relaciones lineales significativas entre el radio del exoplaneta y otras variables del conjunto de datos.

La primera variable con una fuerte correlación es el **cociente del radio del planeta sobre el radio de su estrella anfitriona (pl_ratdor)**. La segunda variable con alta correlación es la **profundidad del tránsito (pl_trandep)**. La tercera variable significativa es la **amplitud de la velocidad radial del planeta (pl_rvamp)**.

Con base en estos resultados, se seleccionan estas tres variables como **variables predictoras** para el radio del exoplaneta, ya que presentan una **alta correlación lineal**, positiva o negativa, con la variable faltante.

Detección de valores atípicos (outliers) en las predictoras

Para las variables predictoras seleccionadas (**pl_ratdor**, **pl_trandep**, **pl_rvamp**), se detectan valores atípicos en el dataset que puedan afectar el análisis y la predicción de la variable faltante (**pl_radj**).



El **boxplot** revela la presencia de **valores atípicos** en las variables predictoras seleccionadas, lo que puede influir significativamente en el análisis y en la precisión de la estimación de la variable faltante. Estos **outliers** pueden sesgar los resultados y reducir la eficacia de los modelos de predicción, especialmente si se utilizan técnicas sensibles a valores extremos, como la regresión lineal estándar.

Para abordar este problema, es recomendable emplear **métodos de regresión robustos**, diseñados para minimizar el impacto de valores atípicos en el ajuste del modelo.

Análisis de regresión

Para explorar la relación entre las características conocidas de los exoplanetas, se lleva a cabo un **análisis de regresión lineal**. El objetivo es predecir el **radio del exoplaneta (pl_radj)** utilizando tres variables que podrían estar estrechamente relacionadas con su tamaño:

- **pl_ratdor**: el cociente entre el radio del exoplaneta y el radio de su estrella, que proporciona una medida relativa del tamaño del planeta.
- **pl_trandep**: la profundidad del tránsito, que indica cuánto disminuye el brillo de la estrella cuando el exoplaneta pasa frente a ella, lo que está directamente relacionado con su radio.
- **pl_rvamp**: la amplitud de la velocidad radial del planeta, que refleja su efecto gravitacional sobre la estrella y puede aportar información sobre su masa y, en combinación con otras variables, su tamaño.

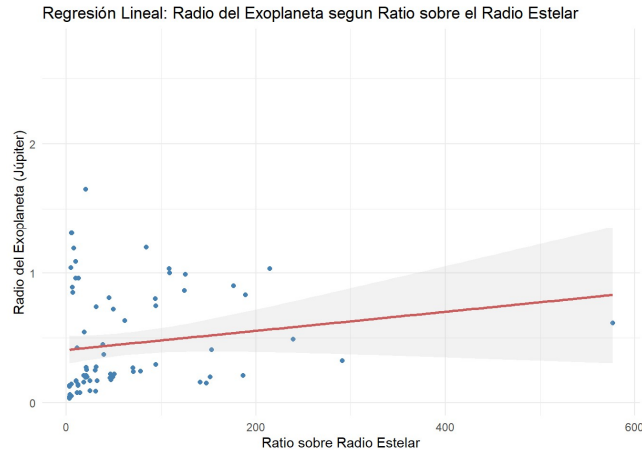
A través de este análisis, se busca comprender mejor cómo estas variables influyen en el tamaño de los exoplanetas y determinar cuál de ellas tiene un mayor impacto en la predicción de **pl_radj**.

Regresión lineal univariada

Se explorará un análisis de regresión lineal univariada con varias variables independientes evaluando la capacidad de predicción de cada modelo planteado.

Regresión lineal: radio del exoplaneta según pl_ratdor

```
## lm(formula = pl_radj ~ pl_ratdor, data = kepler_data_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3749 -0.2695 -0.2125  0.3126  1.2264
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  0.4086542   0.0535610    7.630 0.0000000000489 ***
## pl_ratdor    0.0007341   0.0005069    1.448    0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3872 on 78 degrees of freedom
## (2467 observations deleted due to missingness)
## Multiple R-squared:  0.02618,    Adjusted R-squared:  0.0137
## F-statistic: 2.097 on 1 and 78 DF,  p-value: 0.1516
```



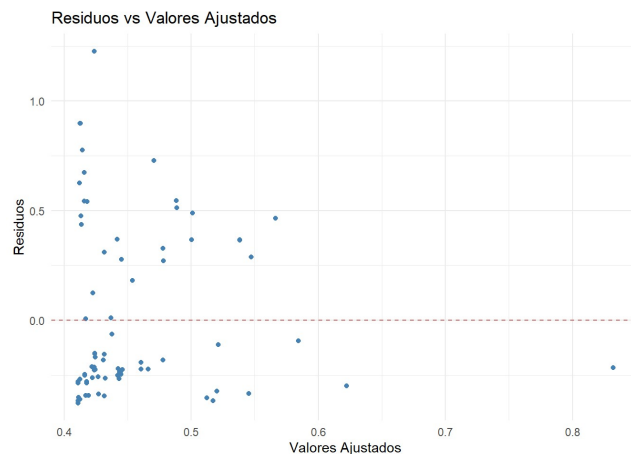
$$pl_radj = \beta_0 + \beta_1 \cdot pl_ratdor + \varepsilon$$

El análisis de regresión indica que **pl_ratdor** no es un buen predictor del radio del exoplaneta (**pl_radj**). El coeficiente de regresión es muy pequeño (0.0007341), y el valor p (0.152) es mayor que 0.05, lo que sugiere que no hay suficiente evidencia estadística para afirmar una relación significativa entre ambas variables.

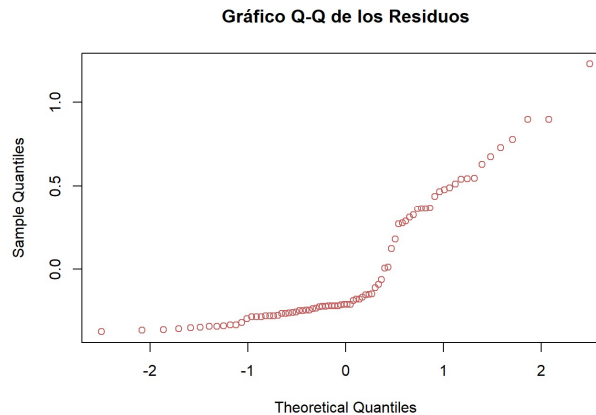
Las estadísticas de ajuste refuerzan esta conclusión. El **Error Estándar Residual** (0.3872) indica una gran desviación entre los valores predichos y reales. El **R² múltiple** (0.02618) muestra que solo el 2.6% de la variabilidad en **pl_radj** es explicada por **pl_ratdor**, y el **R² ajustado** (0.0137) confirma la baja capacidad predictiva del modelo. Además, el **F-statistic** (2.097) y su valor p (0.1516) indican que el modelo completo no es estadísticamente significativo, lo que confirma que **pl_ratdor** no contribuye de manera relevante a la estimación del radio del exoplaneta.

El análisis de los residuos muestra que estos se concentran en los valores ajustados más bajos (0.4-0.5) y tienden a dispersarse a medida que los valores ajustados aumentan. No se observa un patrón curvo, lo que sugiere una relación aproximadamente lineal entre **pl_ratdor** y **pl_radj**.

Dado el bajo **R² (~2.6%)**, era esperable encontrar alta dispersión en los residuos, lo que confirma que **pl_ratdor** no es un buen predictor de **pl_radj**. La presencia de valores atípicos sugiere que existen otras variables más relevantes que explican mejor la variabilidad en el radio del exoplaneta.



El análisis de normalidad de los residuos muestra desviaciones en los cuantiles extremos, donde los puntos se alejan de la línea teórica, indicando la presencia de colas más pesadas de lo esperado en una distribución normal. Esto sugiere la existencia de valores atípicos o que los residuos no siguen completamente una distribución normal. Dado que la regresión lineal asume normalidad en los residuos para garantizar la validez de los valores p e intervalos de confianza, estas desviaciones plantean dudas sobre el cumplimiento de esta suposición.



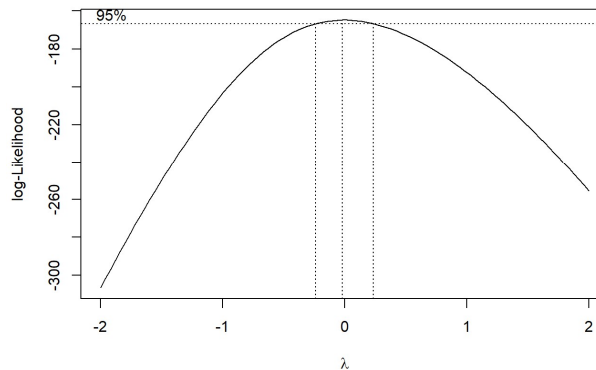
Para mitigar este problema, se podrían aplicar transformaciones en la variable dependiente (**pl_radj**), como logaritmos o raíces cuadradas, para mejorar la normalidad de los residuos.

La prueba de **Shapiro-Wilk** muestra un **estadístico W de 0.81534**, lo que indica una desviación considerable de la normalidad, ya que valores cercanos a 1 sugieren una distribución normal. Además, el **p-valor obtenido (≈ 0.0000001313)** es significativamente menor a 0.05, lo que proporciona evidencia sólida para rechazar la hipótesis nula.

La prueba de **Breusch-Pagan** se utilizó para evaluar si los residuos del modelo presentan homocedasticidad (varianza constante). El **estadístico BP fue 1.1306**, un valor relativamente bajo que no indica una fuerte señal de heterocedasticidad. El **p-value = 0.2876** sugiere que **no se detecta heterocedasticidad significativa en los residuos**, lo que implica que la varianza de los errores es relativamente constante en el modelo.

Para mejorar el modelo, se exploran posibles **transformaciones en la variable dependiente**. Se utiliza la **transformación de Box-Cox** es una técnica utilizada en modelos de regresión para encontrar la mejor manera de transformar la variable dependiente

$$Y' = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}$$



El valor óptimo del parámetro **lambda** obtenido en la transformación de **Box-Cox** es **-0.0202**, que corresponde a una **transformación logarítmica** que podría mejorar la capacidad predictiva. Para verificar si esta transformación realmente mejora el ajuste del modelo, se lleva a cabo una nueva **regresión lineal utilizando la variable transformada**.

```
## lm(formula = pl_adj ~ log(pl_ratdor), data = kepler_data_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3776 -0.2720 -0.2215  0.2916  1.2130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.29440    0.12286   2.396   0.019 *
## log(pl_ratdor)  0.04726    0.03397   1.391   0.168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3876 on 78 degrees of freedom
## (2467 observations deleted due to missingness)
## Multiple R-squared:  0.02421,    Adjusted R-squared:  0.0117
## F-statistic: 1.935 on 1 and 78 DF,  p-value: 0.1681
```

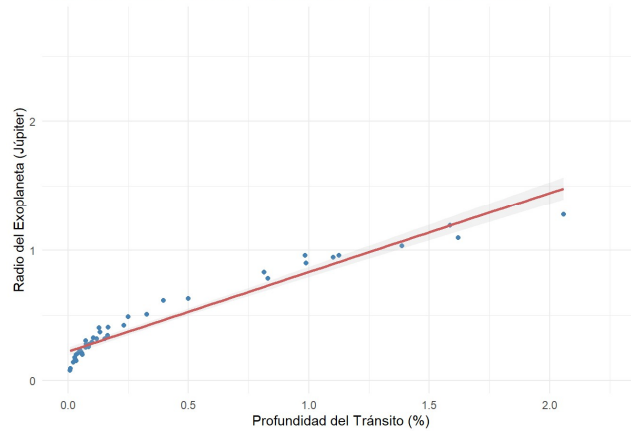
El modelo ajustado con la transformación logarítmica de **pl_ratdor** muestra resultados casi idénticos al modelo original, lo que indica que la transformación **no mejoró significativamente la capacidad predictiva**. El **error estándar de los residuos (0.3876)** y el **R² múltiple (2.4%)** siguen siendo muy bajos, lo que demuestra que la variable transformada apenas explica la variabilidad de **pl_adj**. Además, el **estadístico F (1.935)** y su **p-valor (0.1681)** confirman que no hay suficiente evidencia estadística para considerar que la transformación aporta mejoras. En conclusión, el modelo sigue sin ser adecuado para predecir **pl_adj** a partir de **pl_ratdor**, por lo que se explorarán otras variables o combinaciones de predictores para mejorar su ajuste.

Regresión lineal: radio del exoplaneta según pl_trandep

```
## lm(formula = pl_adj ~ pl_trandep, data = kepler_data_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.203530 -0.056704  0.002386  0.056979  0.150580
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
```

```
## (Intercept) 0.22308 0.01590 14.03 <0.000000000000002 ***
## pl_trandep 0.61058 0.02434 25.08 <0.000000000000002 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08315 on 40 degrees of freedom
## (2505 observations deleted due to missingness)
## Multiple R-squared: 0.9402, Adjusted R-squared: 0.9387
## F-statistic: 629.1 on 1 and 40 DF, p-value: < 0.0000000000000022
```

Regresión Lineal: Radio del Exoplaneta vs Profundidad del Tránsito

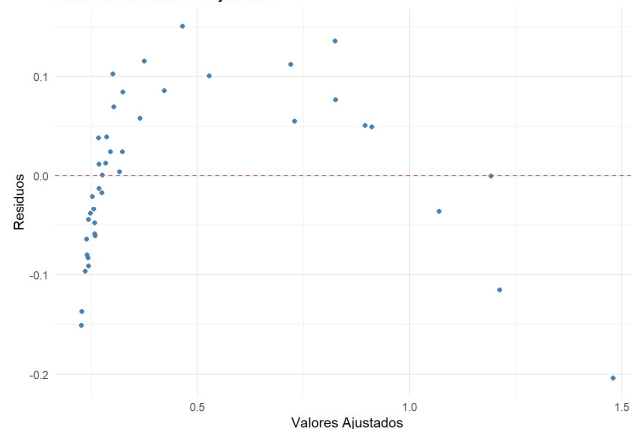


$$pl_radj = \beta_0 + \beta_1 \cdot pl_trandep + \varepsilon$$

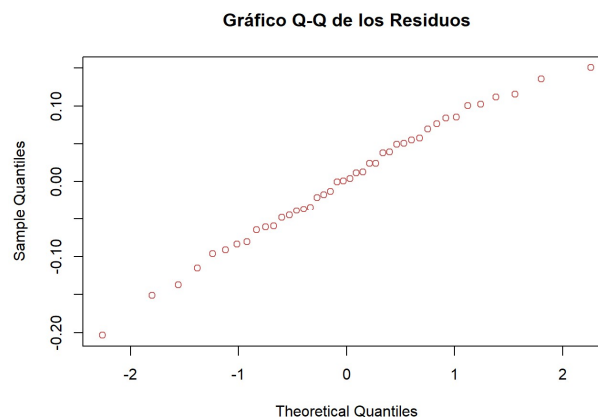
El modelo presenta un **error estándar de los residuos de 0.08315**, indicando una baja variabilidad en las predicciones y un ajuste preciso. El **coeficiente de determinación R^2 es del 94.02%**, lo que significa que el modelo explica casi toda la variabilidad de **pl_radj**. Incluso el **R^2 ajustado (93.87%)** sigue siendo muy alto, confirmando la solidez del ajuste.

La **prueba de significancia global**, evaluada con el **estadístico F (629.1)** y un **p-valor extremadamente bajo**, refuerza la validez del modelo y su capacidad para explicar la relación entre las variables. Comparado con modelos anteriores (**pl_radj ~ pl_ratdor** y **pl_radj ~ log(pl_ratdor)**), este nuevo modelo muestra una mejora drástica en el ajuste. La relación entre **pl_trandep** y **pl_radj** es fuerte y estadísticamente significativa, lo que indica que **pl_trandep es un excelente predictor del radio del exoplaneta**.

Residuos vs Valores Ajustados



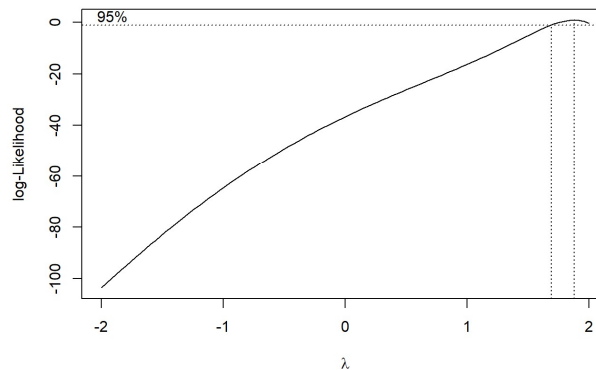
El **gráfico de residuos vs valores ajustados** muestra un **patrón de dispersión desigual**, donde los residuos están más agrupados cerca de **0** en los valores ajustados bajos, pero se dispersan más a medida que estos aumentan. Esto sugiere **heterocedasticidad**, ya que los residuos siguen un **patrón de parábola invertida**, lo que indica que la **varianza de los errores no es constante**. Esto puede afectar la **fiabilidad de las inferencias estadísticas** del modelo, por lo que se recomienda realizar la prueba **de Breusch-Pagan** para confirmar su presencia. Además, se observa una mayor concentración de puntos en la parte izquierda del gráfico, lo que sugiere que el modelo **no está captando completamente la relación entre pl_radj y $pl_trandep$** . Para mejorar la estabilidad del modelo, podría ser necesario **incluir nuevas variables explicativas o aplicar transformaciones** que ayuden a corregir la variabilidad en los errores.



El **gráfico Q-Q de los residuos** muestra que, en general, los puntos se alinean bien con la **línea diagonal**, lo que indica que los residuos siguen un **comportamiento cercano a la normalidad**. **Pequeñas desviaciones en los extremos** sugieren la presencia de **colas más pesadas o algunos valores atípicos leves**. Aunque **no parecen indicar una violación grave de la normalidad**, se complementa el análisis con **pruebas estadísticas**.

La **prueba de Shapiro-Wilk** se utilizó para evaluar si los residuos del modelo siguen una distribución normal. El resultado arrojó un **estadístico W de 0.98715**, un valor cercano a **1**, lo que indica que los datos se ajustan bien a una distribución normal. Además, el **p-valor obtenido fue de 0.9115**, lo que, al ser mayor que **0.05**, significa que no hay suficiente evidencia estadística para rechazar la **hipótesis nula**. En otras palabras, los residuos **no presentan desviaciones significativas de la normalidad**.

El **test de homocedasticidad de Breusch-Pagan** indica la presencia de **heterocedasticidad significativa**, con un **estadístico BP de 6.2822** y un **p-valor de 0.0122**. Dado que este valor es menor que **0.05**, se **rechaza la hipótesis nula**, lo que confirma que la **varianza de los errores no es constante** en todo el rango de valores ajustados. Esto sugiere que el modelo podría beneficiarse de **transformaciones en las variables** o el uso de **métodos más robustos** para corregir la heterocedasticidad.



El **análisis de Box-Cox** indica que la mejor transformación para la variable dependiente es **elevarla al cuadrado (Y^2)**, ya que el **valor óptimo de lambda es ≈ 2** y se encuentra dentro del intervalo de confianza. Esta transformación ayuda a **estabilizar la varianza de los residuos** y mejorar la **linealidad del modelo**. En contraste, **no se recomienda el logaritmo ($\log(Y)$) ni la raíz cuadrada (\sqrt{Y})**, ya que sus valores de lambda están fuera del rango óptimo y no mejorarían significativamente la normalidad ni la homocedasticidad del modelo.

```
## Call:
## lm(formula = I(pl_radj^2) ~ pl_trandep, data = kepler_data_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.108946 -0.015184 -0.004665  0.010102  0.134871
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  0.010364   0.007744   1.338        0.188
## pl_trandep   0.803698   0.011855  67.795 <0.000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04049 on 40 degrees of freedom
## (2505 observations deleted due to missingness)
## Multiple R-squared:  0.9914, Adjusted R-squared:  0.9912
## F-statistic: 4596 on 1 and 40 DF, p-value: < 0.0000000000000022
```

La **transformación cuadrática** mejoró significativamente la precisión y capacidad predictiva del modelo, reduciendo el **error estándar de los residuos a 0.04049** y logrando un **ajuste excepcional** con un **R^2 de 0.9914**. Incluso al corregir por el número de predictores, el **R^2 ajustado se mantiene alto (0.9912)**, lo que confirma la solidez del modelo. Además, el **estadístico F de 4,596** y un **p-valor extremadamente bajo ($< 2.2 \times 10^{-16}$)** indican que el modelo es altamente significativo. En general, esta transformación no solo mejoró el ajuste con respecto al modelo original, sino que también consolidó una **relación fuerte y estadísticamente significativa** entre **pl_trandep** y **pl_radj²**, proporcionando una representación más precisa de la relación entre las variables.

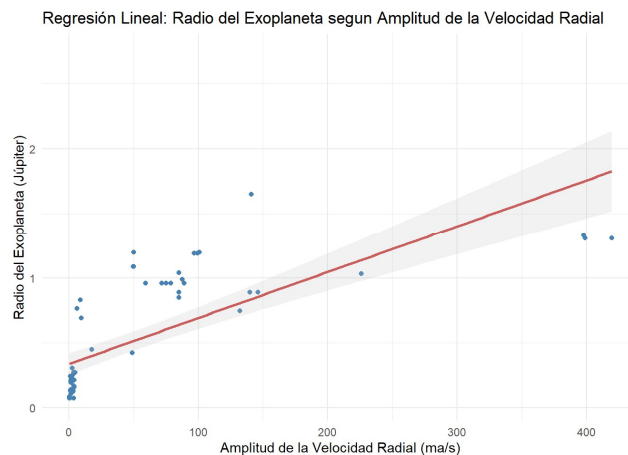
La prueba de **normalidad de Shapiro-Wilk** arroja un **estadístico W de 0.8507** y un **p-valor de 0.00006504**, lo que indica que los residuos **no siguen una distribución normal**. La prueba de **Breusch-Pagan** indica **heterocedasticidad significativa**, con un **estadístico BP de 14.87** y un **p-valor de 0.0001152**, lo que sugiere que la **varianza de los errores no es constante** en todo el modelo.

Aunque la **transformación cuadrática** aplicada a la variable dependiente redujo parcialmente este problema, los resultados muestran que **aún persiste en cierta medida**.

Esto sugiere que, aunque el modelo ha mejorado, puede ser necesario aplicar **otras estrategias**, como regresión ponderada o el uso de errores estándar robustos, para garantizar una estimación más estable y confiable.

Regresión lineal: radio del exoplaneta según pl_rvamp

```
## lm(formula = pl_radj ~ pl_rvamp, data = kepler_data_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5172 -0.2119 -0.1206  0.2644  0.8116
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  0.337785   0.041804   8.080 0.0000000001903 ***
## pl_rvamp      0.003550   0.000406   8.744 0.0000000000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3026 on 66 degrees of freedom
## (2479 observations deleted due to missingness)
## Multiple R-squared:  0.5367, Adjusted R-squared:  0.5297
## F-statistic: 76.46 on 1 and 66 DF, p-value: 0.00000000001239
```



$$pl_radj = \beta_0 + \beta_1 \cdot pl_rvamp + \varepsilon$$

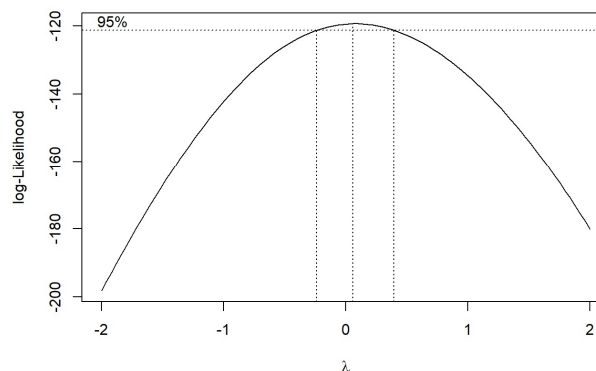
El modelo explica el **53.67% de la variabilidad en pl_radj** ($R^2 = 0.5367$), indicando una relación **moderada a fuerte**, aunque aún queda un **46.33% sin explicar**, lo que sugiere la influencia de otros factores. Incluso al corregir por el número de predictores, el **R^2 ajustado (0.5297)** sigue siendo alto, manteniendo un buen ajuste. Además, el **estadístico F de 76.46** y un **p-valor extremadamente bajo** confirman que el modelo es **altamente significativo**. Aunque su ajuste no es tan preciso como el de $pl_radj \sim pl_trandep$, la relación entre pl_rvamp y pl_radj sigue siendo relevante, lo que indica que **pl_rvamp es un predictor útil**, aunque no el más fuerte disponible.

El resultado de la **prueba de normalidad de Shapiro-Wilk** indica que los residuos **no siguen una distribución normal**, con un **estadístico W de 0.86708** y un **p-valor de 0.000003207**. Dado que este p-valor es extremadamente bajo, hay suficiente evidencia para **rechazar la hipótesis de normalidad**. El resultado del **test de Breusch-Pagan** indica la presencia de **heterocedasticidad significativa**, con un **estadístico BP de 10.096** y un **p-valor de 0.001486**. Esto significa que la

varianza de los errores no es constante, lo que puede comprometer la precisión de las estimaciones del modelo y las pruebas de hipótesis.

Dado que el modelo no cumple con la **suposición de homocedasticidad** y, además, los residuos **no siguen una distribución normal**, se busca corregir la heterocedasticidad mediante una **transformación de la variable dependiente**.

Para determinar cuál es la mejor transformación, se lleva a cabo una **prueba de Box-Cox**, que permite identificar la forma más adecuada para estabilizar la varianza de los errores y mejorar el ajuste del modelo.

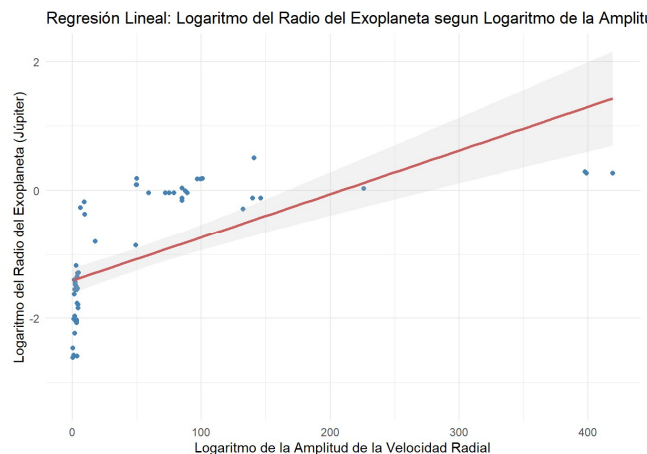


El **pico de la curva** en la prueba de **Box-Cox** se encuentra cerca de **lambda = 0**, lo que indica que la mejor transformación para la variable dependiente se encuentra en esa zona. En términos prácticos, cuando el valor óptimo de **lambda (λ)** es **0**, la transformación recomendada es el **logaritmo natural** de la variable dependiente.

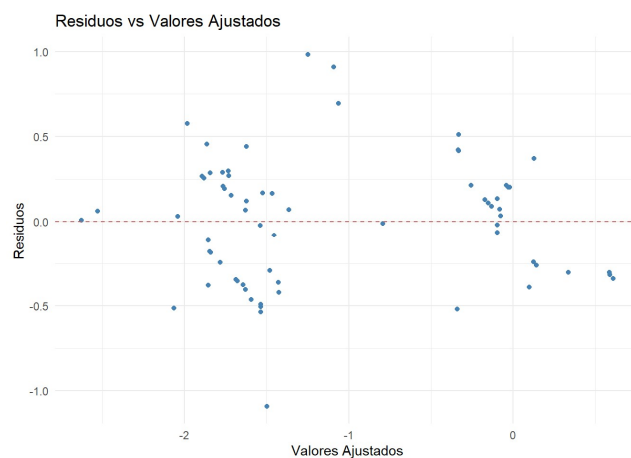
Dado que $\lambda = 0$ está **dentro del intervalo de confianza**, esto confirma que aplicar una **transformación logarítmica ($\log(Y)$)** es la opción más adecuada. Este tipo de transformación es útil porque puede **reducir la heterocedasticidad**, mejorar la **normalidad de los residuos** y hacer que la relación entre las variables sea **más lineal**, lo que permite que el modelo de regresión tenga un mejor ajuste y sea más preciso en sus estimaciones.

```
## lm(formula = log(pl_radj) ~ log(pl_rvamp), data = kepler_data_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09181 -0.30754  0.04893  0.21513  0.98300
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -2.06303     0.06935  -29.75 <0.00000000000002 ***
## log(pl_rvamp)  0.44267     0.02356   18.79 <0.00000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3753 on 66 degrees of freedom
## (2479 observations deleted due to missingness)
## Multiple R-squared:  0.8425, Adjusted R-squared:  0.8401
## F-statistic: 353 on 1 and 66 DF, p-value: < 0.000000000000022
```

La **transformación logarítmica** mejoró significativamente la capacidad predictiva del modelo, aumentando el **R² a 84.25%**, lo que indica un **ajuste excelente** en comparación con el **53.67% del modelo sin transformación**. Incluso al ajustar por el número de predictores, el **R² ajustado se mantiene alto (84.01%)**, confirmando la solidez del modelo. Además, el **estadístico F de 353 y un p-valor extremadamente bajo** muestran que la relación entre las variables es altamente significativa. En conclusión, la transformación logarítmica hizo que el modelo fuera **más preciso y adecuado** para describir la relación entre **pl_rvamp y log(pl_radj)**, proporcionando estimaciones más confiables.

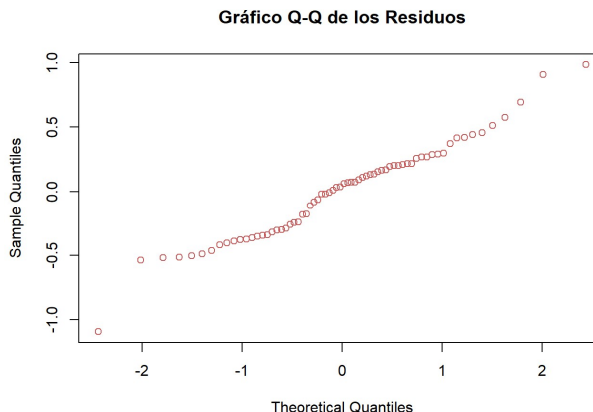


$$\log(pl_radj) = \beta_0 + \beta_1 \cdot pl_rvamp + \varepsilon$$



La prueba de normalidad de Shapiro-Wilk arroja un **estadístico W de 0.97622** y un **p-valor de 0.2188**, lo que indica que no hay evidencia suficiente para rechazar la hipótesis de normalidad. Esto significa que los **residuos siguen una distribución normal**. Si bien el **gráfico Q-Q** ya mostraba una buena alineación de los puntos con la línea diagonal, la prueba estadística **confirma de manera más rigurosa** que los residuos no presentan desviaciones significativas respecto a una distribución normal. Esto refuerza la validez del modelo y la fiabilidad de sus estimaciones.

La prueba de **homocedasticidad de Breusch-Pagan** muestra un **estadístico BP de 0.48865** y un **p-valor de 0.4845**, lo que indica que no hay evidencia de **heterocedasticidad significativa** en el modelo. Esto significa que la **varianza de los errores es constante**, cumpliendo con una de las suposiciones clave de la regresión lineal.



Dado que los modelos anteriores presentaban problemas de heterocedasticidad, estos resultados confirman que **la transformación logarítmica aplicada a la variable dependiente corrigió eficazmente este problema**.

Tras evaluar los **tres modelos de regresión lineal univariada**, se concluye que el modelo **pl_radj ~ pl_rvamp con transformación logarítmica** es el que ofrece el **mejor ajuste**.

$$\log(pl_radj) = \beta_0 + \beta_1 \cdot pl_rvamp + \varepsilon$$

Este modelo explica el **84.25% de la variabilidad en log(pl_radj)**, lo que indica una capacidad predictiva excelente. Además, la relación entre **pl_rvamp y log(pl_radj)** es **fuerte y altamente significativa**, con residuos que siguen una **distribución normal** y **no presentan heterocedasticidad significativa**.

Regresión lineal multivariada

Para mejorar la precisión en la predicción del **radio del exoplaneta (pl_radj)**, se realiza un **análisis de regresión lineal multivariada**, considerando simultáneamente tres variables explicativas: **el cociente del radio planetario sobre el radio estelar (pl_ratdor)**, que mide el tamaño relativo del exoplaneta; **la profundidad del tránsito (pl_trandep)**, relacionada con la disminución de brillo estelar y el tamaño del planeta; y **la amplitud de la velocidad radial (pl_rvamp)**, que aporta información sobre su masa y radio. Este enfoque permite analizar la influencia conjunta de estas variables, en lugar de evaluarlas por separado, logrando un **modelo más preciso y robusto**, capaz de capturar mejor la **variabilidad en los datos** y mejorar las estimaciones del tamaño de los exoplanetas.

```
## Call:
## lm(formula = pl_radj ~ pl_ratdor + pl_trandep + pl_rvamp, data =
kepler_data_num)
##
## Residuals:
##      611      1785      1822      1946      1947      2251      2465      2541
##  0.00903 -0.07158 -0.03219  0.08415 -0.01535  0.09444 -0.08736  0.01885
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 0.2110174 0.0531204 3.972 0.016505 *
## pl_ratdor 0.0001583 0.0004277 0.370 0.730111
## pl_trandep 0.7120088 0.0781077 9.116 0.000803 ***
## pl_rvamp -0.0008151 0.0006926 -1.177 0.304472
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08727 on 4 degrees of freedom
## (2539 observations deleted due to missingness)
## Multiple R-squared: 0.9731, Adjusted R-squared: 0.953
## F-statistic: 48.28 on 3 and 4 DF, p-value: 0.001342
```

El modelo de regresión presenta un **excelente ajuste**, explicando el **97.31% de la variabilidad en pl_{radj}** , con un **R^2 ajustado del 95.3%**, lo que indica una gran capacidad predictiva. Además, el **estadístico F de 48.28 y un p-valor de 0.0013** confirman su alta significancia.

Sin embargo, su confiabilidad se ve comprometida por la **drástica reducción de datos**, ya que solo **4 observaciones efectivas** quedaron tras eliminar valores faltantes, mientras que **2,539 registros fueron descartados**. Esto podría implicar que el modelo está sobreajustado. Para mejorar la estabilidad del modelo, sería necesario **recuperar más observaciones** o simplificar la regresión, eliminando **pl_{ratdor} y pl_{rvamp}** , y dejando solo **$pl_{trandep}$** como predictor, lo que podría mantener un buen ajuste con una estructura más robusta, pero hacer esta eliminación es **volver al segundo modelo analizado**.

Regresión logística

Para predecir la **habitabilidad de los exoplanetas**, se aplica un **modelo de regresión logística**, utilizando como variables independientes los **parámetros planetarios y estelares conocidos**, y como variable dependiente la **habitabilidad**. Dado que se estudian estrellas con radios similares al solar ($\pm 10\%$), se consideran **tres variables clave**:

1. **Temperatura de equilibrio (pl_{eqt})**, fundamental para definir la **zona habitable**, ya que determina la posibilidad de agua líquida en la superficie de un planeta. Kopparapu y otros ([Kopparapu, y otros, 2013](#)) analizan los factores que ajustan los límites de la zona habitable y establecen rangos de temperatura ideales.
2. **Insolación (pl_{insol})**, que mide el **flujo de radiación recibido** por el planeta en comparación con la Tierra, afectando su balance energético. Shields y otros ([Shields, Ballard, & Johnson, 2016](#)) estudian cómo este flujo influye en la zona habitable y en la capacidad de los planetas para retener agua.
3. **Semieje mayor orbital ($pl_{orbsmax}$)**, que define la **distancia del planeta a su estrella** y, por lo tanto, influye en su temperatura superficial. Kane y Gelino ([Kane & Gelino, 2012](#)) demuestran cómo la distancia orbital y la luminosidad de la estrella afectan la habitabilidad.

Determinar la **habitabilidad** es un desafío **multifactorial**, ya que depende de diversos parámetros físicos y químicos. Aunque no existe una fórmula universalmente aceptada, un enfoque común es combinar estos indicadores en una **fórmula ponderada**, considerando la presencia de **agua líquida** como factor esencial. Un ejemplo de esta metodología es el **Earth Similarity Index (ESI)**, propuesto por Méndez y otros ([Méndez, y otros, 2021](#)), que utiliza una combinación normalizada de **temperatura, flujo estelar y parámetros orbitales** para estimar el potencial de habitabilidad de un exoplaneta.

Así con los indicadores seleccionados definimos un **Índice de Habitabilidad Planetaria (PHI - Planetary Habitability Index)** como una función normalizada que combine los indicadores clave:

$$PHI = w_1 \cdot f_1(pl_eqt) + w_2 \cdot f_2(pl_insol) + w_3 \cdot f_3(pl_orbsmax)$$

Donde:

- w_i son los pesos normalizados asociados a cada indicador sumando 1.
- $f_i(x)$ son funciones de transformación o normalización que convierten las medidas en valores entre 0 y 1, según el rango de habitabilidad conocido.

Estas variables se consideran como posibles predictores de la habitabilidad de los exoplanetas cuando están en determinados rangos:

- **Temperatura de equilibrio (pl_eqt):** Idealmente debe estar en un rango compatible con agua líquida, aproximadamente entre 273K (0 °C) a 373K (100 °C). Sin embargo, debido a factores atmosféricos y de presión, este rango puede extenderse ligeramente: $T_{min} = 200K$ considerando atmósferas densas como la de Marte y $T_{max} = 400K$ considerando atmósferas con alta presión como la de Venus. Según la literatura especializada el peso de la temperatura de equilibrio en el índice de habitabilidad es de $w_1 = 0.40$ cuando se considera la temperatura de equilibrio como un indicador preponderante de habitabilidad y no se tienen muchos más datos del planeta

$$f_1(pl_eqt) = \max\left(0, \min\left(1, \frac{pl_eqt - T_{min}}{T_{max} - T_{min}}\right)\right)$$

- **La insolación (pl_insol):** la cantidad de energía recibida por el planeta se mide en unidades de flujo de energía recibida por unidad de área y se puede normalizar en un rango de 0 a 1. La insolación ideal para la vida es aquella que permite la presencia de agua líquida en la superficie del planeta: $1S$ (donde S es la insolación solar a la distancia de la Tierra). La vida puede existir en un rango de insolación de $I_{min} = 0.3S$ (límite interior de la zona habitable para el Sol, más el agua podría congelarse) a $I_{max} = 1.7S$ (límite exterior de la zona habitable, más allá el agua podría evaporarse). Según la literatura especializada el peso de la insolación en el índice de habitabilidad es de $w_2 = 0.30$.

$$f_2(pl_insol) = \max\left(0, \min\left(1, \frac{pl_insol - I_{min}}{I_{max} - I_{min}}\right)\right)$$

- **La distancia orbital del semieje mayor (pl_orbsmax):** el valor ideal para la vida es aquella que permite la presencia de agua líquida en la superficie del planeta; La Tierra tiene una distancia orbital semieje mayor de 1 UA. La vida puede existir en un rango de $D_{min} = 0.5$ UA (límite interior de la zona habitable para el Sol, más allá el agua podría evaporarse) a $D_{max} = 2$ UA (límite exterior de la zona habitable, más allá el agua podría congelarse). Se tomará un peso de $w_3 = 0.30$.

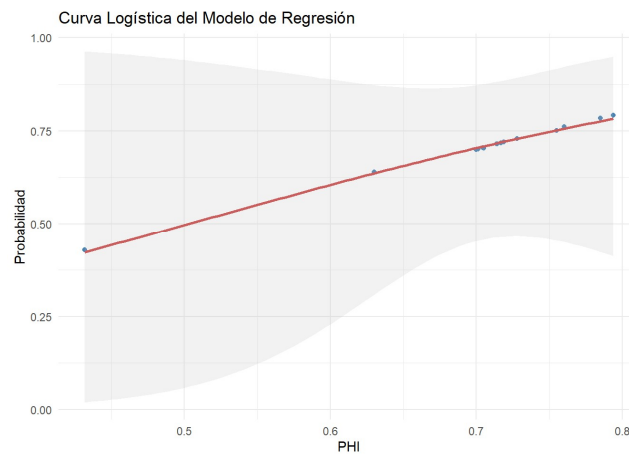
$$f_3(pl_orbsmax) = \max\left(0, \min\left(1, \frac{pl_orbsmax - D_{min}}{D_{max} - D_{min}}\right)\right)$$

La fórmula para calcular el **PHI** que se utilizará en el modelo de regresión logística es:

$$PHI = 0.4 \cdot f_1(pl_eqt) + 0.3 \cdot f_2(pl_insol) + 0.3 \cdot f_3(pl_orbsmax)$$

Se calculan las variables normalizadas con los límites descritos y se calcula PHI con los pesos explicados en un nuevo dataset para encarar la regresión logística.

```
## Call:
## glm(formula = phi ~ ., family = binomial, data =
kepler_data_logistic_normalized)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.488      7.878  -0.316   0.752
## pl_eqt_normalized    2.228    11.813   0.189   0.850
## pl_insol_normalized    1.105     6.760   0.163   0.870
## pl_orbsmax_normalized    1.569     6.123   0.256   0.798
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 0.45354364  on 16  degrees of freedom
## Residual deviance: 0.00063461  on 13  degrees of freedom
## (2530 observations deleted due to missingness)
## AIC: 19.641
##
## Number of Fisher Scoring iterations: 4
```



$$\log\left(\frac{P(\phi = 1)}{1 - P(\phi = 1)}\right) = \beta_0 + \beta_1 \cdot \text{pl_eqt_normalized} + \beta_2 \cdot \text{pl_insol_normalized} + \beta_3 \cdot \text{pl_orbsmax_normalized}$$

El **modelo de regresión logística** obtenido no es sólido debido a la **escasez de datos disponibles** para calcular el **PHI (Planetary Habitability Index)**. Con solo **17 observaciones completas**, la **baja cantidad de datos** limita la **precisión y confiabilidad** del modelo, afectando su capacidad de **generalización** y estabilidad en los coeficientes. Para mejorar su validez, sería necesario contar con un **mayor número de exoplanetas con datos completos**, lo que permitiría entrenar el modelo con más información y obtener **predicciones más precisas sobre habitabilidad**.

El modelo presenta **sobreajuste**, evidenciado por la **drástica reducción en la desviación residual** de **0.4535 a 0.0006**, lo que indica que se ajusta demasiado a los **pocos datos disponibles**, afectando su capacidad de generalización. Aunque el **AIC (19.641)** sugiere una penalización adecuada por complejidad, la **falta de significancia en los coeficientes** indica que el modelo **no es útil para la predicción**. Además, la **eliminación de 2,530 observaciones por datos faltantes**, dejando solo **17**

registros, genera **grados de libertad muy bajos ($df = 13$)**, afectando la estabilidad y confiabilidad de las estimaciones.

En su estado actual, el modelo no es útil, ya que **no cuenta con variables predictoras significativas** y enfrenta **una pérdida masiva de datos**, lo que limita su capacidad para hacer inferencias válidas. Antes de sacar conclusiones sobre la relación entre **PHI** y las variables predictoras, sería recomendable **limpiar los datos y probar un modelo más simple**. Sin embargo, dada la **escasez de observaciones**, es posible que no sea factible construir un modelo confiable para la **predicción de habitabilidad** en estas condiciones.

Posibles mejoras para futuros análisis

Para mejorar el **modelo de regresión logística**, una posible solución es abordar el problema de los **valores faltantes** en las variables clave utilizadas para calcular el **PHI**: **temperatura de equilibrio (pl_eqt)**, **insolación (pl_insol)** y **semieje mayor orbital ($pl_orbsmax$)**. En lugar de descartar todas las observaciones con datos faltantes, se puede aplicar **regresión lineal** para estimar estos valores en función de otras variables relacionadas dentro del conjunto de datos. Al hacer esto, se logra recuperar una mayor cantidad de observaciones, lo que aumentaría el tamaño de la muestra utilizada en la regresión logística y mejoraría la estabilidad del modelo.

Sin embargo, utilizar variables estimadas mediante regresión como **input en otro modelo predictivo** conlleva ciertos **riesgos e implicaciones estadísticas**. Cuando una variable ha sido generada mediante una predicción previa, **su error de estimación se transfiere al modelo final**, en este caso, a la regresión logística. Como resultado, el modelo puede **subestimar la incertidumbre real** y producir **intervalos de confianza demasiado optimistas**, ya que no está considerando los errores asociados a las predicciones previas.

Además, el uso de **valores estimados por regresión** puede introducir **colinealidad artificial**, especialmente si la variable imputada depende de otras variables que ya están incluidas en la regresión logística. Esto puede distorsionar los coeficientes y hacer que el modelo sea menos interpretable. También es posible que la **estructura de los errores en las variables imputadas** no sea idéntica a la de los valores observados originalmente, lo que podría generar un sesgo en las estimaciones del modelo.

Para mitigar estos riesgos, se pueden aplicar estrategias como la **validación cruzada** para evaluar el impacto de las variables imputadas en el desempeño del modelo y realizar un **análisis de sensibilidad**, probando el modelo con y sin las variables imputadas para verificar la estabilidad de los resultados. Además, una opción más robusta sería utilizar **métodos de imputación múltiple**, que generan varias versiones de los datos imputados y permiten incorporar la incertidumbre en las predicciones finales.

Conclusiones

El análisis realizado permitió evaluar diferentes modelos de regresión para predecir el radio de los exoplanetas y explorar la relación entre sus características físicas. Se probaron modelos de regresión lineal y logística, incluyendo transformaciones para mejorar la normalidad de los residuos y corregir problemas de heterocedasticidad. Los resultados indicaron que el modelo **$pl_radj \sim pl_rvamp$ con transformación logarítmica** fue el más efectivo, explicando el **84.25% de la variabilidad** en el radio del exoplaneta y cumpliendo con los principales supuestos estadísticos. Sin embargo, el modelo de **regresión logística** para predecir la habitabilidad planetaria presentó limitaciones debido a la escasez de datos completos, lo que afectó su estabilidad y capacidad predictiva.

Para mejorar la calidad de los modelos, es fundamental abordar el problema de los valores faltantes, ya sea mediante **imputación de datos** o la recolección de más observaciones. Además, al utilizar variables estimadas en modelos posteriores, se debe considerar el impacto de los **errores de predicción acumulados** en los resultados finales. Si bien las transformaciones de datos ayudaron a mejorar la precisión de algunos modelos, la falta de datos sigue siendo una limitación clave para la predicción de la **habitabilidad planetaria**. Futuras investigaciones deberían centrarse en ampliar el conjunto de datos y explorar enfoques más robustos, como modelos de aprendizaje automático, que podrían mejorar la capacidad predictiva de estos análisis.

Referencias

- Alonso Sobrino, R. (2006). *Detección y Caracterización de Exoplanetas Mediante El Método de Los Tránsitos*. Instituto de Astrofísica de Canarias (IAC). Obtenido de <https://www.iac.es/es/ciencia-y-tecnologia/publicaciones/deteccion-y-caracterizacion-de-exoplanetas-mediante-el-metodo-de-los-transitos>
- Brachman, R. Z. (2024). *How to Analyze Your Data. How to Get Started*. Obtenido de Exoplanet Exploration: Planets Beyond Our Solar System: <https://exoplanets.nasa.gov/exoplanet-watch/how-to-contribute/how-to-analyze-your-data/>
- Cambazard, H., Catusse, N., Chomez, A., & Lagrange, A.-M. (2025). Logistic regression to boost exoplanet detection performances. *Monthly Notices of the Royal Astronomical Society*, Volume 536, Numero 2, Páginas 1610–1624. doi: <https://doi.org/10.1093/mnras/stae2657>
- Cardenas, C., Lozano, D., Marquez, C., Torres, E., & Delgado-Correal, C. (2022). Optimización de un sistema difuso para la detección automática de tránsitos planetarios en curvas de luz de estrellas individuales. *Ciencia en Desarrollo*, 1. 19-35. doi: <http://dx.doi.org/10.19053/01217488.v1.n2E.2022.15136>
- Cermak, A. (2024). *Kepler / K2*. Obtenido de NASA Science: <https://science.nasa.gov/mission/kepler/>
- Dong-Hong, W. (2023). The possibility of detecting our solar system through astrometry. *Research in Astronomy and Astrophysics*. doi: <https://doi.org/10.48550/arXiv.2309.11729>
- IPAC. (2021). *NASA Exoplanet Archive Overview and Holdings*. Obtenido de NASA Exoplanet Science Institute: <https://exoplanetarchive.ipac.caltech.edu/docs/intro.html>
- IPAC. (2024). *Planetary Systems and Planetary Systems Composite Parameters Data Column Definitions*. Obtenido de NASA Exoplanet Science Institute: https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html
- IPAC. (2025). *NASA Exoplanet Archive*. Obtenido de NASA Exoplanet Science Institute: <https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS>
- IPAC. (s.f.). *Our Mission*. Obtenido de IPAC Out Mission: <https://www.ipac.caltech.edu/page/mission>
- Kane, S. R., & Gelino, D. M. (2012). *The Habitable Zone and Extreme Planetary Orbits*. doi: <https://doi.org/10.1089/ast.2011.0798>
- Kopparapu, R. K., R. R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., . . . Deshpande, R. (2013). *Habitable Zones Around Main-Sequence Stars: New Estimates*. doi: <https://doi.org/10.1088/0004-637x/765/2/131>
- Malik, A. M., & Obermeier, C. (2021). Exoplanet detection using machine learning. *Monthly Notices of the Royal Astronomical Society*. doi: <https://doi.org/10.1093/mnras/stab3692>
- Marin, D. (2011). *Detectando Planetas Desde El Espacio Gracias a Einstein*. Obtenido de Eureka: <https://danielmarin.naukas.com/2011/05/23/detectando-planetas-desde-el-espacio-gracias-a-einstein/>
- Marin, D. (2018). *EarthFinder: Un Telescopio Espacial Para Buscar Exotierras Por El Método de La Velocidad Radial*. Obtenido de Eureka: <https://danielmarin.naukas.com/2018/03/14/earthfinder/>

- Méndez, A., Rivera-Valentin, E., Schulze-Makuch, D., Filiberto, J., Ramirez, R., Wood, T., . . . Byrne, P. (2021). *Habitability Models for Astrobiology*. doi: <https://doi.org/10.1089/ast.2020.2342>
- Nardi, L. (2024). *¿Cuáles Son Las Técnicas Utilizadas Para Descubrir Exoplanetas?* Obtenido de WIRED: <https://es.wired.com/articulos/cuales-son-las-tecnicas-utilizadas-para-descubrir-exoplanetas>
- NASA. (2022). *Webb de La NASA Obtiene Su Primera Imagen Directa de Un Mundo Distante*. Obtenido de NASA Ciencia: <https://ciencia.nasa.gov/universo/webb-de-la-nasa-obtiene-su-primera-imagen-directa-de-un-mundo-distante/>
- Nicolau, J. E. (2025). *Ra-Tp-Final*. Obtenido de Github: <https://github.com/georgsmeinung/ra-tp-final/>
- Pimentel, J., Amorim, J., & Rudzicz, F. (2024). Feature extraction for exoplanet detection. *International Journal of Data Science and Analytics*. doi: <https://doi.org/10.1007/s41060-024-00552-7>
- Shields, A. L., Ballard, S., & Johnson, J. A. (2016). The habitability of planets orbiting M-dwarf stars. *Physics Reports*, Volume 663, Páginas 1-38. doi: <https://doi.org/10.1016/j.physrep.2016.10.003>
- Tukey, J. W. (1977). *Exploratory Data Analysis* (18 ed., Vol. 2). (A.-W. P. Company, Ed.)
- Venkata, S. R., Jahnavi, B. C., Venkata, R. R., & Muvva, S. (2023). Exoplanet Detection Using Feature Engineering with Ensemble Learning. *3rd International Conference on Pervasive Computing and Social Networking ICPCSN 2023*, 116-122. doi: <http://dx.doi.org/10.1109/ICPCSN58827.2023.00025>
- Wolszczan, A., & Frail, D. A. (1992). A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, Volume 355, Páginas 145-147. doi: <https://doi.org/10.1038/355145a0>