Statistical Analysis of Kepler third law using Open Exoplanet Catalogue

Aliaksandr Hubin

July 31, 2024

1. Research Question

Derive physical laws using statistical techniques based on the Open Exoplanet Catalogue Tables (Rein 2012). I suggest to concentrate on the 3rd Kepler's law with the response variable `semimajoraxis` (see Table 3). I suggest working with no prior assumption of the underlying true law, however, assume that the law exists and can be discovered using these data.

2. Details and Background

The Open Exoplanet Catalogue is a database containing information on all discovered extra-solar planets. The repository provides simple ASCII tables derived from the original XML files, capturing planetary system details. The data can be read using the following R command

```
   data = read.csv(”https://raw.githubusercontent.com/OpenExoplanetCatalogue/oec_tables/
master/comma_separated/open_exoplanet_catalogue.txt”)
```

3. Dataset and Relevant Variables

The dataset consists of extra-solar planets, including information on their identifiers, orbital parameters, and host star details. Overall summary of the dataset is provided in Table 1.

Name	Value
Rows	5,414
Columns	25
Discrete columns	7
Continuous columns	18
All missing columns	0
Missing observations	46,203
Total observations	135,350

Table 1: Basic Statistics - Raw Counts

Name: The primary identifier of the planet serves as a unique label, facilitating cross-referencing with other databases and publications. It is analogous to a celestial social security number.

Mass: The planetary mass, expressed in Jupiter masses, plays a pivotal role in gravitational interactions within the planetary system. It influences the strength of gravitational forces, shaping orbits and determining the planet's overall dynamics.

Radius: The planetary radius, measured in Jupiter radii, provides insights into the physical size and composition of the planet. This parameter is crucial for understanding the potential habitability and structure of the celestial body.

Orbital Period: The orbital period, measured in days, is fundamental to Kepler's laws. It defines the time it takes for a planet to complete one full orbit around its host star, influencing the dynamics of the planetary system.

Semi-major Axis: Expressed in Astronomical Units (AU), the semi-major axis defines the size of the planet's orbit. This parameter indicates the average distance between the planet and its host star, providing valuable information about the planet's position in the system and the potential habitable zone.

Eccentricity: Eccentricity characterizes the shape of the planet's orbit. A value of 0 represents a perfectly circular orbit, while higher values indicate more elongated, elliptical paths.

Periastron Angle: The periastron angle, specified in degrees, identifies the point in the orbit where the planet is closest to its host star. This angle contributes to understanding the asymmetry in the planet's motion and its interaction with the host star.

Orbital Inclination: Orbital inclination, measured in degrees, describes the tilt of the planet's orbit relative to the reference plane. It provides insights into the orientation of the planet's path and its potential impact on climate and axial tilt.

Surface Temperature: The surface or equilibrium temperature, given in Kelvin, is crucial for assessing the potential habitability of the planet. It may influence atmospheric conditions, surface temperature, and the presence of liquid water.

Discovery Method: Understanding the method used to discover the planet is essential for evaluating potential biases and limitations in the dataset. Different discovery methods, such as radial velocity or transit observations, contribute to the dataset's diversity.

Discovery Year: The discovery year provides temporal context to the dataset, allowing researchers to analyze trends and advancements in exoplanet discovery techniques over time.

Host Star Properties: Parameters such as host star mass, radius, metallicity, temperature, and age significantly impact the planetary system. The host star's characteristics influence the luminosity received by the planets, the composition of planetary atmospheres, and the overall stability of the system.

Introductory EDA

Density plots of the most continious covariates, which are of primary interest are presented in Figure 1, while the correlations between these covariates are depicted in Figure 2. Further, density of log of eccentricity, mass, period, semimajoraxis, hoststar_radius, hoststar_mass are plotted in Figure 3.

By considering these parameters, researchers can unravel the intricate dynamics of planetary systems, draw connections between physical properties and observed phenomena, and contribute to a deeper understanding of our celestial neighbors. Each field encapsulates a unique aspect of the exoplanetary realm, allowing for comprehensive analyses and meaningful interpretations within the realm of astrophysics.

4. Estimand

The estimand focuses on deriving Kepler's 3rd law from the given dataset, exploring the relationships between planetary motion and their respective star systems. The true law in Newton's formulation is

$$\text{semimajoraxis} \propto \left((\text{hoststar\_mass} + \text{mass}) \times \text{period}^2\right)^{1/3}$$

which can (due to mass $\ll$ hoststar_mass) be approximated by Kepler's formulation of the law

$$\text{semimajoraxis} \propto \left(\text{hoststar\_mass} \times \text{period}^2\right)^{1/3}$$

Indeed, running linear regresssion assuming the Keplers formulation of the law gives the result presented in Table 2. **But how can we derive the law without knowing apriori its closed form solution?**

Variable	Estimate	Std. Error	t value	Pr(> |t|)
I(((hoststar_mass) * (period^2))^(1/3))	2.337e-02	7.899e-05	295.8	<2e-16 ***

Table 2: Linear regression summary with *Multiple R-squared:* 0.972, *Adjusted R-squared:* 0.972

5. Proposed Model and Modeling Approach

The proposed model involves statistical techniques to analyze the orbital parameters and identify patterns that align with Kepler's 3rd law. I propose to build a regression with `semimajoraxis` used as response and a subset of other reasonable variables from Table 3 as the covariates.

6. Predictive Modeling with a Descriptive Goal

The analysis aims to predict planetary semi major axis based on the available data, with a descriptive goal to understand and validate Kepler's 3rd law.

7. Statistical Issues Addressed in this Task

Key statistical challenges include handling missing data, performing model selection and interpeting its parameters. We assume to be in an *M-Closed* setting.

Field	**Description**
name	Primary identifier of the planet
binaryflag	Binary flag [0=no known stellar binary companion; 1=P-type binary (circumbinary); 2=S-type binary; 3=orphan planet (no star)]
mass	Planetary mass [Jupiter masses]
radius	Radius [Jupiter radii]
period	Period [days]
semimajoraxis	**Semi-major axis [Astronomical Units]**
eccentricity	Eccentricity
periastron	Periastron [degree]
longitude	Longitude [degree]
ascendingnode	Ascending node [degree]
inclination	Inclination [degree]
temperature	Surface or equilibrium temperature [K]
age	Age [Gyr]
discoverymethod	Discovery method
discoveryyear	Discovery year [yyyy]
lastupdate	Last updated [yy/mm/dd]
system_rightascension	Right ascension [hh mm ss]
system_declination	Declination [+/-dd mm ss]
system_distance	Distance from Sun [parsec]
hoststar_mass	Host star mass [Solar masses]
hoststar_radius	Host star radius [Solar radii]
hoststar_metallicity	Host star metallicity [log relative to solar]
hoststar_temperature	Host star temperature [K]
hoststar_age	Host star age [Gyr]
list	A list of lists the planet is on

Table 3: Fields Description in the Open Exoplanet Catalogue

References

Rein, H. (2012), 'A proposal for community driven and decentralized astronomical databases and the open exoplanet catalogue', *arXiv preprint arXiv:1211.7121* .

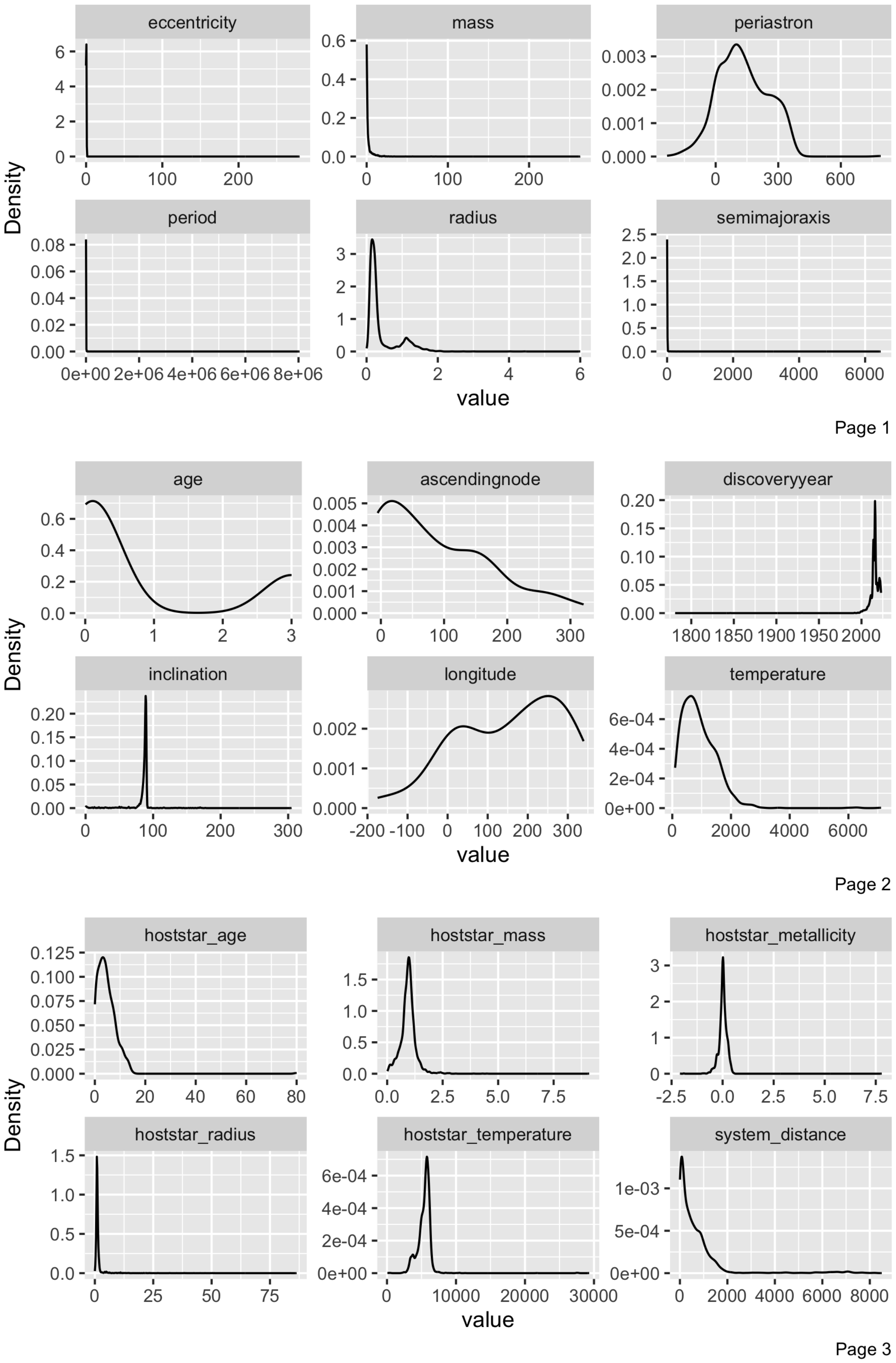

Figure 1: Density plots of the continous covariates

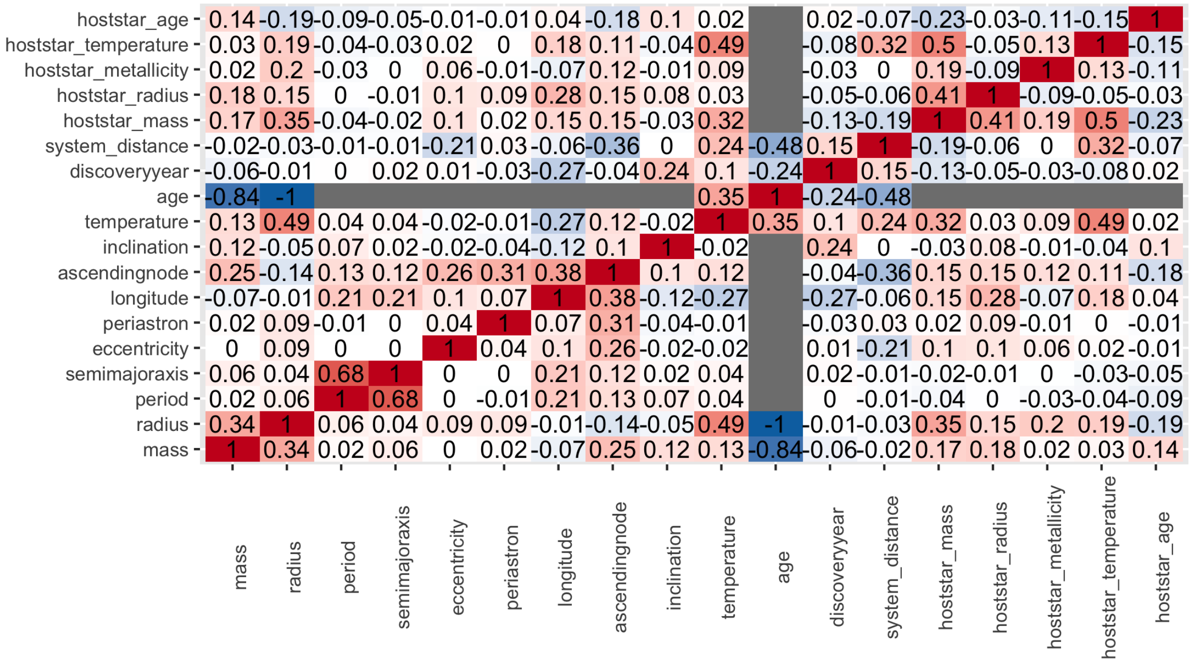

Figure 2: Pearson correlations between the continous covariates

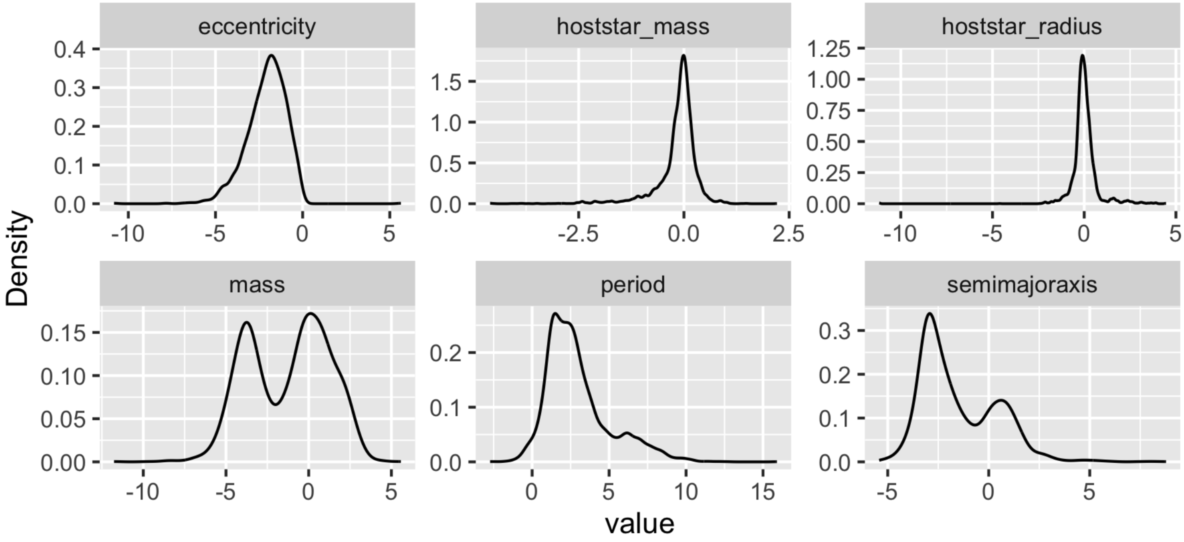

Figure 3: Density plots of the logorithm of selected continous covariates