THE INFLUENTIAL EFFECT OF BLENDING, BUMP, CHANGING PERIOD, AND ECLIPSING CEPHEIDS ON THE LEAVITT LAW

A. García-Varela, J. R. Muñoz, B. E. Sabogal, S. Vargas Domínguez, and J. Martínez

1 Universidad de los Andes, Departamento de Física, Cra. 1 No. 18A-10, Bloque Ip, Bogotá, Colombia; josegarc@uniandes.co.edu.co, jmuñoz2198@uniandes.edu.co, bsabogal@uniandes.edu.co
2 Universidad Nacional de Colombia—Sede Bogotá—Facultad de Ciencias—Observatorio Astronómico—Carrera 45 No. 26-85, Bogotá, Colombia; svargasd@unal.edu.co
3 Universidad de los Andes, Departamento de Ingeniería Industrial, Edificio ML, Cra 1 Este No 19A-40, Bogotá, Colombia; j.martinez144@uniandes.edu.co

Received 2015 August 28; accepted 2016 April 14; published 2016 June 14

ABSTRACT

The investigation of the nonlinearity of the Leavitt law (LL) is a topic that began more than seven decades ago, when some of the studies in this field found that the LL has a break at about 10 days. The goal of this work is to investigate a possible statistical cause of this nonlinearity. By applying linear regressions to OGLE-II and OGLE-IV data, we find that to obtain the LL by using linear regression, robust techniques to deal with influential points and/or outliers are needed instead of the ordinary least-squares regression traditionally used. In particular, by using M- and MM-regressions we establish firmly and without doubt the linearity of the LL in the Large Magellanic Cloud, without rejecting or excluding Cepheid data from the analysis. This implies that light curves of Cepheids suggesting blending, bumps, eclipses, or period changes do not affect the LL for this galaxy. For the Small Magellanic Cloud, when including Cepheids of this kind, it is not possible to find an adequate model, probably because of the geometry of the galaxy. In that case, a possible influence of these stars could exist.

Key words: Magellanic Clouds – methods: statistical – stars: variables: Cepheids

Supporting material: machine-readable table

I. INTRODUCTION

By studying variable stars in the Small Magellanic Cloud (SMC), Henrietta Leavitt discovered a linear relation between the pulsation period and magnitudes of 25 Cepheids, in the sense that brighter Cepheids have longer periods (Leavitt & Pickering 1912). This correlation, commonly called the period–luminosity (PL) relation, was later renamed the Leavitt law (LL) in honor of its discoverer (Freedman & Madore 2010).

This linear statistical correlation is a cornerstone in the measurement of extragalactic distances using stellar standard candles. However, in order to obtain an accurate calibration of this relation, it is important to establish the effects of nonlinearities, metallicity, and companions on it, as mentioned by Madore & Freedman (1991), as well as considering other effects such as Cepheids showing the Hertzsprung progression and Cepheids exhibiting changes in period. A brief literature review of these issues is presented below.

(i) The nonlinearity effect. The first works in this direction began 25 years after the discovery by Henrietta Leavitt. In a study of the shape of the PL relation of nearby galaxies, Kukarkin found that this relation has a break around 10 days (Fernie 1969). Three decades later, Sandage & Tammann (1968, 1969) found evidence of curvature in the PL relation, and afterward, Tammann et al. (2003) confirmed the break in the PL relation at 10 days. Kanbur & Ngeow (2004) again confirmed this result based on the statistical F-test. In recent years, the following studies have confirmed the nonlinearity of the LL without proposing an explanation for it: the testimator method (Kanbur et al. 2007), the approaches of linear regression residuals and additive models (Koen et al. 2007), the multiphase PL relations (Ngeow et al. 2012), and the multiple least-squares regression (García-Varela et al. 2013).

(ii) The metallicity effect. By using nonlinear convective pulsating models, Caputo et al. (2000) and Marconi et al. (2005) studied synthetic multiband PL relations of populations of Cepheids of different chemical compositions, uniformly distributed over the instability Strip. The masses of these stars follow the law $dn/dm = m^{-3}$, and are distributed in the range $5\text{–}11 M_\odot$. For a wide range of log $P$, those authors found clear evidence that the optical LL is better represented by a quadratic relationship than by a linear one, showing a dependence on the metallicity and the intrinsic width of the instability strip. For IR bands they showed that the LL is better fitted by a linear function showing a slightly dependence on the metallicity.

UDalski et al. (2001) and Pietrzyński et al. (2004) found strong evidence in favor of the universality of the optical PL relation, in the metallicity range from $-1.0$ to $-0.3$ dex. Gieren et al. (2005) and Fouqué et al. (2007) found that the slopes of the PL relations in $V/W_0$ do not change significantly between the environments of the Milky Way and the LMC.

(iii) The companion effect. Physically bound companions to Cepheids are difficult to detect at distances of tens of kiloparsecs. To produce a detectable effect on the light curve, the eclipses should be deep enough and a significant number of points associated with the eclipses should be observed. The first condition is achieved with an adequate combination of radius, luminosity, and effective temperature of the primary and secondary stars, jointly with the appropriate inclination and mass ratio of the system. These eclipsing systems exhibit light curves similar to those of Cepheids, showing in addition dispersed points below them, which are a signal of the eclipses (UDalski et al. 1999a, Figure 7). When the light curve of the Cepheid is subtracted, a typical eclipsing signal emerges (Pietrzyński et al. 2008, Figure 3).

Nun et al. (2014) claim to detect a few Cepheids in eclipsing systems in the Galactic Bulge and the Magellanic Clouds, using a random-forest supervised algorithm over the MACHO catalog of variable stars. However, because these variables do...
not show a clear period of eclipses, it is difficult to affirm that they are binaries. It is very likely that these stars are blended with near neighbors. Soszyński et al. (2015) reported a few Cepheids in the Magellanic Clouds that could be blended, and also others that could be members of binary systems.

Detection of Cepheid systems whose light curves exhibit eclipsing variations is also important: if the Cepheid is a member of a physical binary system, it is possible to determine dynamical masses, radii, and distance with very high accuracy. Until now, spectroscopic studies of double-lined eclipsing binary systems in the LMC have been made by the Optical Gravitational Lensing Experiment (OGLE) in the systems OGLE-LMC-CEP-02274 (Pietrzyński et al. 2010), OGLE-LMC-CEP-1812 (Pietrzyński et al. 2011), OGLE-LMC-CEP-1718 (Gieren et al. 2014), OGLE-LMC-CEP-2532 (Pilecki et al. 2015), and OGLE-LMC562.05.9009 (Soszyński et al. 2012; Gieren et al. 2015)—a system detected in the OGLE Gaia South Ecliptic Pole Field.

(iv) The Hertzsprung progression effect. Inspecting light curves of 37 Cepheids, Hertzsprung (1926) found a relation between the position of a bump feature in the light curve and the pulsation period. Later studies confirmed that this bump feature is present in the light and radial curves of fundamental-mode classical Cepheids whose periods are in the range 6–16 days (Bono et al. 2000). This feature is present in the descending branch for pulsations in the range 6–9 days. For longer periods, this feature is observed in the ascending branch, and it disappears for periods longer than 20 days (Bono et al. 2000; Gastine & Dintrans 2008). The amplitude bump grows when the period increases, reaching its maximum value for periods near to the maximum light around 10–11 days. For longer periods, the amplitude bump decreases until it vanishes (Keller & Wood 2006).

There are two models that explain some of the properties of the observed bump feature. The first one, called the echo mechanism, proposes radial-pressure waves generated in the He ii ionization region (Christy 1968). A wave traveling inward is reflected in the core and reaches the surface one period later, leading to the formation of the bump (Bono et al. 2000). However, this model has two points to solve: the contradiction with the acoustic-ray formalism (Whitney 1983; Aikawa & Whitney 1984, 1985) and the difficulty of predicting adequately the Christy wave velocity near to the stellar surface (Karp 1975; Bono et al. 2000).

The second model, known as the resonance mechanism, proposes that the bump feature arises from a resonance between the fundamental mode and the second overtone (Simon & Schmidt 1976). By using numerical simulations Gastine & Dintrans (2008) studied the nonlinear saturation of the acoustic modes excited by the $\kappa$-mechanism. They found that this 2:1 resonance causes the bump to appear in the ascending branch for $P_2/P_0 < 1/2$.

(v) The changing-period effect. Long-term observations have detected several thousand oscillation cycles for short-period Cepheids. This allows changes in period to be determined with very high accuracy. For long-period Cepheids, it is not possible to reach this accuracy, because the data in the worst cases of sampling contain only tens of cycles. There are three scenarios that explain different characteristics of the rates of change in period. The first one proposes that the evolutionary changes in the stellar structure of Cepheids crossing the instability strip are responsible for the rates of change in period (Turner et al. 2006; Fadeyev 2013). The second one proposes that the presence of magnetic fields is a possible explanation for the randomly changing periods exhibited by some Cepheids. The last one suggests that the amplitude and phase variations of Cepheids are caused by an effect analogous to the Blazhko effect, as exhibited by RR Lyrae stars (Molnár & Szabados 2014). Soszyński et al. (2015) reported a few Cepheids that could exhibit this Blazhko effect, the number of these Cepheids being larger in the LMC than in the SMC.

The changes in period should produce a shift in the phase of the light curve, and consequently a higher scatter in the light curve, detectable after years of observations. A study of 655 LMC Cepheids found changes in period in 18% of the fundamental mode, and in 41% of the first overtone (Poleski 2008, Table 3). A visual inspection of the light curves of these OGLE Cepheids shows, for most of them, a strong scatter of the points as a consequence of the shift in phase.

Since there are few studies about magnetic fields on cool, radially pulsating stars, it is not clear whether the nature of these magnetic fields is fossil-like or whether they are produced as a consequence of the stellar pulsation (Wade et al. 2002). Based on a solar-like magneto-convective cycle, Stothers (2009) could explain the observed rates of change in period of two short-period Cepheids: Polaris and V473 Lyr. As the existence of local magnetic fields on the surface of late-type stars causes spots (Strassmeier 2009), Neilson & Ignace (2014) proposed that convective hot spots can be a possible explanation for the random changes in the pulsation period detected in the single Cepheid observed by the Kepler mission, V1154 Cyg.

The first detection of magnetic fields in variable stars was made on RR Lyrae (Babcock 1958). Measurements of magnetic fields for Cepheid stars began later on. In particular, for $\alpha$ Car and $\gamma$ Cyg there were reported values of 700 G (Weiss 1986) and from ~100 to ~350 G (Severny et al. 1974; Plachinda 1990), respectively. Spectropolarimetric works on the bright bump Cepheid $\eta$ Aql ($V = 3.90, P = 7.17$ days) reported controversial results. While Plachinda (2000) and Butkovskaya et al. (2014) measured periodic variations of the longitudinal magnetic field with an amplitude of tens of gauss, Wade et al. (2002) found an insignificant detection of the longitudinal magnetic field at a level of 10 G.

In order to establish whether or not the effects previously described are influential on the LL, we make a study using statistical techniques. We begin, in the second section, with a brief description of the LMC LL and the OGLE-II and OGLE-IV Cepheid data. The third section is dedicated to reviewing the statistical theory of linear regression analysis, relevant for this work. The fourth section presents the results of applying statistical models in order to obtain an optical LL of the LMC and SMC. Finally, our main conclusions are given in the fifth section.

2. LMC LL WITH OGLE DATA

OGLE-II and OGLE-IV observations of Cepheid variables in the LMC and SMC galaxies were collected with the 1.3 m Warsaw telescope, at Las Campanas Observatory, Chile (Udalski et al. 1999a, 1999b, 2015). While Cepheid catalogs for the OGLE-II fundamental mode contain 771 and 1319 stars for the LMC and SMC, respectively, OGLE-IV has a nearly
complete collection (2429 and 2739 for the LMC and SMC, respectively), covering practically the whole Magellanic System with a time baseline of a little more than five years (Soszyński et al. 2015).

The slope and zero point of the LMC LL in optical VI-bands were computed by Udalski (2000) using ordinary least-squares (OLS) regression, on OGLE-II fundamental-mode Cepheids. Points deviating by more than $2.5\sigma$ (outliers) were removed by applying the sigma-clipping algorithm (Udalski et al. 1999c). Udalski realized that the standard deviation of residuals of the LMC data was almost half that of the SMC. This greater dispersion is caused mainly by the spatial distribution of Cepheids in the SMC bar, whose thickness is placed along the line of sight with a typical depth of $\sim 0.25$ mag (Harris & Zaritsky 2006). These facts, and the manner in which the LL of the LMC is much better populated for periods longer than 2.5 days, persuaded Udalski to adopt as the universal slope value the one obtained for the LMC, notwithstanding that the number of Cepheids in the SMC was around twice the number in the LMC. For OGLE-IV LMC Cepheids the slope and zero point of the LL were determined in an analogous approach, but the SMC slope was obtained independently, using only Cepheids in this galaxy (Soszyński et al. 2015).

With the aim of establishing a possible statistical cause of the nonlinearity of the LL, we use the dereddened sample of OGLE-II fundamental-mode Cepheids belonging to the LMC and SMC galaxies reported by Udalski et al. (1999a, 1999b), as well as the nearly complete collection of OGLE-IV Cepheids in the Magellanic Clouds reported by Soszyński et al. (2015).

3. LINEAR REGRESSION ANALYSIS

In this section we summarize the topics of the statistical theory of linear regression analysis that are most relevant to this paper.

We start by explaining what the outliers and influential points are. Then, we present the theory of the OLS method and the conditions necessary to apply it. Next, we describe the statistical tools to identify influential points. Then, we present the specification test, which verifies whether all variables involved in a statistical problem are adequately included and represented by a proposed model. Following this, we present the structural break test, which looks for unstable parameters. Finally, we present the robust $M$-regression that allows a linear regression of data with non-Gaussian error distribution. The MM-regression is applied in order to get a high breakdown point (BDP) and fit the models in the presence of influential points that are outliers at the same time.

An extended statistical discussion of these topics can be found in Montgomery et al. (2012). Some applications in the context of astronomy can be found in Feigelson & Babu (2012).

3.1. OLS Statistical Theory

The most common method to estimate parameters in a linear regression model is OLS, based on the Gauss–Markov theorem. The model is (in matrix notation)

$$y = X\beta + \varepsilon,$$  

where $y$ is the dependent variable vector, $X$ is the independent variables matrix, $\beta$ is the parameters vector, and $\varepsilon$ is the errors vector.

This theorem states that the OLS estimators are the best linear unbiased estimators if the expected value of the errors vector is zero ($E(\varepsilon) = 0$), the variance of the errors is constant ($V(\varepsilon) = \sigma^2I$), and the errors are uncorrelated. It can be shown (Montgomery et al. 2012) that OLS estimators ($\hat{\beta}$) are obtained from

$$\hat{\beta} = (X'X)^{-1}X'y.$$  

Notice that $0$ is the null vector, $I$ is the identity matrix, and $A'$ indicates the transpose of matrix $A$.

Once the model is fitted its adequacy must be checked. That is: (i) the error term $\varepsilon$ has a constant variance $\sigma^2I$. (ii) The errors are uncorrelated. These items are the assumptions of the Gauss–Markov theorem. (iii) The errors are normally distributed. This is very important for testing the hypothesis and estimating confidence intervals. (iv) All the observations have approximately the same weight. This refers to the fact that it is not desirable that model estimators depend more on a few observations than on the majority of them. This could happen if some observations, called influential points, have a disproportionate impact in the OLS estimations. (v) There are no specification problems in the model and there is no structural break, i.e., the parameters are stable. This refers to the correct functional form of the model.

After fitting the model, a residual analysis is made to verify its adequacy. Since the residuals are the differences between the original observed values and their fits, they also measure the variability of the dependent variable that is not explained by the model. Therefore, any violations of the assumptions can be detected by analyzing model residuals. Mainly the analysis on residuals is looking for evidence of: (i) errors that come from a distribution with heavier tails than normal. Large departures from normality in the error distribution mean that the $F$- and $t$-tests are no longer valid, and neither are the estimations of the confidence interval. (ii) Heteroskedasticity, meaning that the errors do not have constant variance. (iii) Influential points. (iv) Specification problems. (v) Autocorrelated errors. In this study it is not necessary to check whether errors are correlated because the data are not time-dependent, i.e., they come from a cross-section in the statistical sense.

A normal probability quantil–quantil (Q–Q) plot is used for checking the error distribution. If all the residuals lie along a line, it means that they come from a normal distribution. For more details see Montgomery et al. (2012). This plot shows whether the distribution has heavier or lighter tails or whether it is skewed compared to the normal distribution.

To check the assumption that errors have constant variance (homoskedasticity) White’s test (White 1980) is used in the following equation:

$$y = X\hat{\beta} + \varepsilon,$$  

where $\varepsilon$ is the residual from the linear regression defined as

$$\hat{\varepsilon} = \varepsilon = y - \hat{y}.$$  

This test does not make the assumption that errors come from a normal distribution. Its null hypothesis is that $\sigma^2 = \sigma^2_i$ for all $i$. Since the PL relation can be modeled as a simple linear regression, the test uses the auxiliary model given by

$$e_i^2 = \alpha_0 + \alpha_1x_i + \alpha_2x_i^2 + u_i, \quad i = 1, 2, \ldots, n,$$  

where $u_i$ is the error term and $n$ is the number of stars. It can be shown that White’s statistic $nR^2$ is asymptotically distributed
as $\chi^2$ with $g - 1$ degrees of freedom, where $n$ is the number of observations and $R^2$ is the regression coefficient of determination of Equation (5). Because the auxiliary model (Equation (5)) has three parameters ($g = 3$), $\chi^2$ has two degrees of freedom.

### 3.2. Outliers and Influential Points

Outliers are points that have an unusual behavior. Let us imagine a scatter plot: outliers are data points that unusually are located out of the pattern or far from the data cloud. On the other hand, influential points are those data that have a disproportionate impact on OLS estimators (slope and zero point). As a result, the model estimators depend more on them than on the majority of data. Let us imagine the scatter plot again: a regression line is fitted to all points. Then we delete a point and fit the regression line once more. If the regression line obtained after removing the point changes a lot, then the deleted observation is an influential point. What is a lot? Statistics helps us in this decision, as we shall describe later.

Now, let us imagine the scatter diagram with a point that is distant from the cloud of data points but located along the regression line. If we delete it, the regression line will not change. Therefore, the point is an outlier but not an influential point, because it affects the $R^2$ coefficient and the OLS estimator standard errors, but not the OLS regression estimators. In contrast, if the outlier affects the estimators of the regression coefficient it is also an influential point. Finally, let us imagine that the point is not an outlier; however, after deleting it, the new fitted regression line is very different from that obtained with it. This means that it is an influential point. In general, influential points may or may not be outliers.

**Detection of influential points:** in order to look for influential points Cook’s distance, $DFFITS$, $DFBETAS$, and $COVRATIO$ statistics are commonly used (Montgomery et al. 2012). The first of them is defined as follows:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' \mathbf{X} \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})}{p \hat{\sigma}^2}, \quad i = 1, 2, ..., n. \quad (6)$$

Basically, it measures the square distance between the OLS estimate $\hat{\beta}$ based on all observations and the OLS estimate $\hat{\beta}_{(i)}$ obtained after deleting the $i$th observation, where $p$ in Equation (6) is the number of parameters in Equation (3), i.e., the number of independent variables plus the zero point. Moreover, $\hat{\sigma}^2$ is an estimate of the mean square error, defined as follows:

$$\hat{\sigma}^2 = \frac{\mathbf{e}' \mathbf{e}}{n - p}. \quad (7)$$

Observations with large Cook’s distance affect OLS estimates of $\beta$. To know what is a large Cook’s distance, $D_i$ is compared to the $F_{0.5(p,n-p)}$ distribution (Montgomery et al. 2012). If the $i$th observation has $D_i > F_{0.5(p,n-p)}$, it is considered an influential point.

$DFFITS$ measures how large the influence is of the $i$th observation on its own estimated $\hat{y}_i$. In other words, it shows how many standard deviations $\hat{y}_i$ changes due to the deletion of the $i$th observation. It is defined as follows:

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{(i)}}{\hat{\sigma}(i) \hat{h}_i}, \quad i = 1, 2, ..., n. \quad (8)$$

$\hat{y}_{(i)}$ and $\hat{\sigma}(i)$ are the fitted values and the standard deviation respectively, obtained without the $i$th observation. The term $\hat{h}_i$ is the $i$th diagonal element of the matrix $H$ that is defined as follows:

$$H = \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'.$$

The term $\hat{h}_i$ is the amount of leverage that the $i$th observation has on the $j$th fitted value. A point is considered influential when $|DFFITS| > 2\sqrt{\frac{p}{n}}$. For complementary explanations and references about $DFFITS$ and the $H$ matrix see Montgomery (2012) and Draper & Smith (1998).

Another statistic that detects influential points is $DFBETAS$. It indicates how much effect the $i$th observation has on each $\beta_j$, measured in units of the standard deviation. In order to apply it, the coefficient estimators of $\beta_j$ obtained using all observations are compared with the coefficient estimators computed excluding the $i$th observation ($\hat{\beta}_{(i)}$). As a result, a measure for each $\beta_j$ is obtained. The $DFBETAS$ statistics is defined as follows:

$$DFBETAS_{\hat{\beta}_j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{(i)}(j)}{\hat{\sigma}(i) \hat{C}_{jj}} \quad j = 0, 1, ..., p \quad (10)$$

where $C = (\mathbf{X}' \mathbf{X})^{-1}$, so that $\hat{C}_{jj}$ is the $j$th diagonal element of the matrix $C$ and $\hat{\sigma}(i)$ is the square root of the regression mean square error fitted without the $i$th observation. An observation is considered an influential point if $|DFBETAS_{\hat{\beta}_j(i)}| > 2\sqrt{\frac{p}{n}}$.

The last statistic used to detect influential points is $COVRATIO$, which measures how much the covariance matrix is affected by the $i$th observation. It compares the generalized variance of the parameter estimators obtained without the $i$th observation with that obtained using all observations.

The variance–covariance matrix of parameter estimators with all observations is

$$Var(\hat{\beta}) = \hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1}. \quad (11)$$

The generalized variance is the determinant of Equation (11):

$$GV(\hat{\beta}) = |\hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1}|. \quad (12)$$

Given this, $COVRATIO$ is defined as follows:

$$COVRATIO_i = \frac{|\hat{\sigma}_{(i)}^2 (\mathbf{X}' \mathbf{X}_{(i)})^{-1}|}{|\hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1}|} \quad i = 1, 2, ..., n. \quad (13)$$

The $i$th observation is considered influential if its $COVRATIO$ lies outside the interval $1 - \frac{3p}{n} < COVRATIO_i < 1 + \frac{3p}{n}$, where $p$ is the number of parameters in Equation (3) and $n$ is the total number of observations.

The LL reported in this work is obtained by rejecting from the sample the influential points and Cepheids exhibiting commonalities, as explained later. However, two questions arise that require detailed answers: why is it important to identify these influential points and give them a statistical treatment? Why is it necessary to apply robust techniques to make a linear regression instead of applying a well-known algorithm such as sigma-clipping? In the next paragraphs, we answer these questions, illustrating the associated statistical problem.
Outliers are observations with anomalous behavior. When they are caused by human errors or instrumental failures, they can be recognized and excluded from the analysis. Moreover, they can be rejected from the analysis only if there are strong non-statistical reasons that support such a decision. In this work it is possible to reject them for astronomical reasons. However, the usual practice of deleting outliers, without making a further analysis to look for reasons that could explain their behavior, could lead to serious consequences in the precision of estimators, because they are adjusted artificially (Montgomery et al. 2012).

As a result, when there are no astronomical reasons to exclude those points and/or there are influential points, robust estimation methods are needed.

3.3. Specification Tests

To establish that there is no evidence of specification errors in the model, we use tests to see whether a nonlinear combination of the fitted values is significant to explain the response variable. One of them is Ramsey’s test, given by Equations (14)–(16) (Ramsey 1974). The other one is a variant proposed by Godfrey and Orme that uses only the model given by Equation (14) (Godfrey & Orme 1994). These auxiliary models and their respective hypotheses used in this work are the following:

\[ y = X\hat{\beta} + \alpha_1 y^2 + \epsilon. \] (14)

The null and alternative hypotheses are \( H_0: \alpha_1 = 0 \) and \( H_1: \alpha_1 \neq 0 \)

\[ y = X\hat{\beta} + \alpha_1 y^2 + \alpha_2 y^3 + \epsilon. \] (15)

The null and alternative hypotheses are \( H_0: \alpha_1 = \alpha_2 = 0 \) and \( H_1: \) at least one \( \alpha_i \neq 0; \ i = 1, 2 \)

\[ y = X\hat{\beta} + \alpha_1 y^2 + \alpha_2 y^3 + \alpha_3 y^4 + \epsilon. \] (16)

The null and alternative hypotheses are \( H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0 \) and \( H_1: \) at least one \( \alpha_i \neq 0; \ i = 1, 2, 3. \)

If any of these tests are significant because there is evidence to reject the null hypothesis \( H_0 \), then there is enough evidence of misspecification. In fact, this means that a nonlinear combination of the fitted values is significant enough to explain the variability in the dependent variable, i.e., the functional form is incorrect.

3.4. Structural Breaks

Dummy variables are used to verify whether there are structural breaks. Before explaining what a dummy variable is, it is necessary to give a brief explanation about qualitative variables. Qualitative or categorical variables are those that do not have a natural scale of measurement. Two categorical variables are used in this study. The first variable indicates which data set the observation belongs to when data are divided into two sets, to make structural tests of the models. This variable has two categories: data set 1 and data set 2. The second variable is the name of the galaxy where the Cepheid is located, which also has two categories: LMC and SMC.

Dummy variables are used to indicate which category of a qualitative variable the observations belong to. Each dummy variable takes two possible values: 1 if the observation belongs to the category that the dummy variable represents, and 0 if not. If a qualitative variable has \( K \) categories, it can be represented by \( K \) dummy variables, one for each category.

However, only \( K - 1 \) dummy variables are used in the regression model. Besides, it does not matter which \( K - 1 \) dummy variables are used since the \( H \) matrix is the same independently of which dummy variable is omitted to fit the model. In fact, dummy variables representing the categories of a qualitative variable are linked to each other because when one of them takes a value of 1 for an observation, the others take 0 for the same observation. Therefore, when all the \( K - 1 \) dummy variables in the model take a value of 0, they are representing the dummy variable that is not present in the model. For more details see Montgomery et al. (2012).

The process of verifying the existence of structural breaks is as follows: the data are divided into two data sets by dummy variables. Then, a model that includes the dummy variables is estimated to verify whether the estimated parameters associated with dummies are significant or not. If they are significant, the parameters are unstable, and thus there are structural breaks.

The \( P \)-value is used to contrast hypotheses. This value is the lowest level of significance that leads to rejection of the null hypothesis when it is true. Therefore, if the \( P \)-value is greater than the significance level, there is not enough evidence to reject the null hypothesis. In contrast, if the \( P \)-value is less than or equal to the significance level there is enough evidence to reject the null hypothesis and accept the alternative one, which means the test is significant.

3.5. M-regression

Sometimes OLS assumptions are not satisfied; for example, errors have a distribution with heavier tails than the normal distribution and/or there are influential points. An alternative way to fit a linear model in scenarios of this kind is a robust regression, in which the residuals can be defined as follows:

\[ e_i = y_i - x_i'\beta \quad i = 1, 2, \ldots, n. \] (17)

It is convenient to scale the residuals by using median absolute deviation:

\[ S = \frac{\text{median} |e_i - \text{median}(e_i)|}{0.6745}. \] (18)

Therefore the scaled residuals \( (u_i) \) can be expressed as follows:

\[ u_i = \frac{y_i - x_i'\beta}{S} \quad i = 1, 2, \ldots, n. \] (19)

If scaled residuals are not dependent on each other, and they all have the same distribution \( f(u) \), the maximum likelihood \( \beta \) estimators are those that maximize the likelihood function:

\[ L(\beta) = \prod_{i=1}^{n} f \left( \frac{y_i - x_i'\beta}{S} \right). \] (20)

Draper & Smith (1998) show that this condition is equivalent to minimizing Equation (21). Therefore, a class of robust estimator that minimizes this equation is called an M-estimator and the regression based on it is called M-regression. \( M \) comes from maximum likelihood since the function \( r(u_i) \) is related to Equation (20) for a proper choice of the error distribution.
Besides, \( \rho(u_i) \) weights the scaled residuals \( u_i \) in the sum
\[
\sum_{i=1}^{n} \rho \left( \frac{y_i - x_i' \beta}{S} \right) = \sum_{i=1}^{n} \rho(u_i). \tag{21}
\]
There is more than one suggestion for \( \rho(u_i) \) in the literature; the function used in this work is Tukey’s bi-square, with parameter equal to 6 (Kafadar 1983). For further information about this function and other choices see Draper & Smith (1998). To minimize Equation (21) the first partial derivative of \( \rho(u_i) \) must be calculated:
\[
\frac{\partial \rho}{\partial \beta_j} = \psi \left( \frac{y_i - x_i' \beta}{S} \right) = \psi(u_i) \quad i = 1, 2, \ldots, n, \quad j = 0, 1, 2, \ldots, p. \tag{22}
\]
As a result, the following equation is obtained:
\[
\sum_{i=1}^{n} x_{ij} \psi \left( \frac{y_i - x_i' \beta}{S} \right) = 0, \quad j = 0, 1, 2, \ldots, p \tag{23}
\]
where \( p \) is the number of parameters in the model. Multiplying Equation (23) by \( u_i/u_i \) gives
\[
\sum_{i=1}^{n} x_{ij} w_i \psi \left( \frac{y_i - x_i' \beta}{S} \right) = 0, \quad j = 0, 1, 2, \ldots, p \tag{24}
\]
where \( w_i = \psi(u_i)/u_i \forall y_i = x_i' \beta \), otherwise it is 1.

The determination of the \( M \)-estimator is an iterative process, thus the \( M \)-estimator for \( \beta \) is
\[
\hat{\beta}_{q+1} = (X' W_q X)^{-1} X' W_q y. \tag{25}
\]
This iterative process stops when a convergence criterion is reached, i.e., when the estimators change by less than a preselected amount in the \( (q+1) \)th iteration. Notice that Equation (25) is the weighted least-squares estimator where \( W_q \) is a diagonal \( n \times n \) matrix of weights whose elements are \( w_i \). See Draper & Smith (1998) and Montgomery et al. (2012) for a wider discussion.

### 3.6. MM-regression

In the context of \( M \)-regression it is important to check that there are no bad influential points, i.e., observations that are outliers in the \( X \) space and influential points at the same time. In this scenario the \( MM \)-estimator (Yohai 1987) should be used because it has a high BDP. In a finite sample the BDP is the smallest fraction of contaminated data to cause a divergence of the estimator from the value that it would take if the data were not contaminated (Montgomery et al. 2012).

The \( MM \)-estimator combines the asymptotic efficiency of the \( M \)-estimator with the high BDP estimators such as \( S \)-estimators and Least Trimmed Squares. \( MM \)-regression is obtained in three steps. First, a model for high-BDP parameter estimators is calculated to compute the residuals of the model. Second, based on residuals computed in the previous step, an \( M \)-estimate of scale with high BDP is calculated. Finally, the model parameters are estimated using \( M \)-estimators and the scale estimation computed in the previous step.

Neither normal errors nor homoskedasticity are assumptions in \( M \)-regression or \( MM \)-regression. However, these regressions assume that errors are uncorrelated, which is not an issue when data come from a cross-section, meaning that data are not time-dependent.

Robust regressions are based on asymptotic results. This means they could lead to wrong results when they are used with small data sets. Montgomery et al. (2012) say that a small to moderate sample size could be less than 50 points. In our case, the smallest data set has about 400 observations. Hence, we have enough data to apply these techniques without problems. In particular, OGLE-IV has around twice and three times as many observations as OGLE-II for the SMC and the LMC, respectively.

### 4. LMC AND SMC LL REGRESSIONS

The statistical theories explained previously are applied to both the LMC and SMC galaxies. The following models are fitted by OLS:
\[
V_0 = \beta_1 + \beta_2 LP + \varepsilon \tag{26}
\]
\[
I_0 = \beta_1 + \beta_2 LP + \varepsilon \tag{27}
\]
\[
W_1 = \beta_1 + \beta_2 LP + \varepsilon \tag{28}
\]
where \( LP \) is \( \log_{10} P \) and \( P \) is the pulsation period of a Cepheid. In general, these models violate the OLS assumptions, as will be shown below.

Influential points are data that truly affect the results of linear regression, although they may or may not be outliers. It is very important that studies of the LL treat those points appropriately. Deleting a subset of influential points does not prevent the appearance of new influential points, as will be shown below.

The sigma-clipping algorithm is an iterative procedure commonly used in astronomy in order to reject the outliers in OLS regression. However, to apply the OLS regression its assumptions must be satisfied. In order to investigate whether the OGLE-II data fulfill these conditions, we perform an experiment in the \( V \)-band using the whole sample of 765 OGLE-II LMC fundamental-mode Cepheids. We apply the sigma-clipping algorithm using five iterations and an optimum threshold of 2.5\( \sigma \). These values are the same as used in the studies of the LL by Udalski et al. (1999a) and Garcia-Varela et al. (2013). Simultaneously, we apply White’s test to check whether the assumption of homoskedasticity is fulfilled.

Before the first sigma-clipping iteration, White’s test finds enough evidence to reject the null hypothesis of homoskedasticity with a \( P \)-value of 0.0165 at 5% significance. A total of 79 influential points are detected using some of the Cook’s distance, \( DFFITS, DBETAS, \) and \( COVRATIO \). On the other hand, by applying the criteria of the sigma-clipping algorithm, 27 points are rejected in the first iteration, not all of them identified as influential points. During the second iteration, White’s test again finds enough evidence to reject the null hypothesis of homoskedasticity with a \( P \)-value of 0.0483 at 5% significance, and 76 influential points are identified. By using the sigma-clipping algorithm 22 points are rejected. In the following three iterations influential points are detected again. Taking into account these facts, we notice that the main problems of applying the sigma-clipping algorithm in the context of the LL are:

(i) OLS estimators obtained by using this procedure are not adequate since their precision is adjusted artificially (Montgomery et al. 2012).

(ii) In the two initial iterations, the models have no constant error variance, i.e., they show heteroskedasticity. In the
remaining iterations, the models show homoskedasticity, as a result of deleting points using the sigma-clipping criterion.

(iii) The residuals in each iteration have distributions with heavier tails than the normal distribution.

(iv) Influential points are detected in each iteration.

In summary, rejecting outliers using the sigma-clipping algorithm is not adequate in this case, since there are violations of OLS assumptions. Without strong reasons to support the algorithm is not adequate in this case, since there are violations of OLS assumptions, it is possible to affirm that one cause of nonlinearities of the LL is the inclusion of a fraction of these stars that, despite being Cepheids, are influential points.

To test this hypothesis, we fit the models given by Equations (26)–(28) by OLS using only Cepheids with small dispersion. The results confirm the hypothesis, as will be described below. For this reason, we consider it appropriate to summarize here the communalities of these influential Cepheids:

(i) The scatter exhibited by light curves of some Cepheids classified as A2 or A3 can be associated with phase shifts, suggesting changes in the period of pulsation. This effect produces a dispersion of the mean magnitudes, which moves the locus of the Cepheid in the PL relation.

(ii) The B2 light curves exhibit a behavior that could be explained by changes in period or blending effects. The dispersion of these light curves does not allow us to directly establish which of these effects could be present in these types of Cepheids. In this work it is not clear which of them is more likely to be the best explanation. However, this point is out of the scope of this paper. This interesting fact could be investigated in more detail in future works.

A small fraction of the light curves classified as B3 have characteristics suggesting either eclipses or blending. In the first case, the main effect of a companion on the PL relation is to move the locus of the Cepheid. Because of this, the total mean brightness out of eclipse includes the companion’s contribution, adding scatter in some cases up to 1.0 mag, as was reported by Alcock et al. (2002) and Pietrzyński et al. (2010). The second case is more likely to occur, since a visual inspection of the finding charts of these B3 Cepheids shows that most of them are blended.

(iii) Light curves classified as C1 or C2 are bump Cepheids, i.e., they exhibit the Hertzsprung progression. The amplitude bump affects the Cepheid’s mean magnitude. This effect is stronger for periods near to maximum light, shifting the Cepheid’s position in the PL relation.

In order to use the OLS regression, White’s test is applied. This test does not find enough evidence to reject its null hypothesis of homoskedasticity, with a 5% significance, since the P-values associated with each model test are 0.084, 0.280, and 0.908 for the LMC, and 0.068, 0.154, and 0.440 for the SMC, respectively. However, two issues are noted: the first one is the presence of influential points. The second is that the residuals distributions for V/WI LMC data (and only for the W_I SMC data) have heavier tails than the normal distribution.

Since there is neither an astronomical nor a statistical reason to delete those influential points, or even a fraction of them, a new approach to estimating the parameters given in the models (26)–(28) is needed instead of OLS. Besides, if these influential points were deleted, new ones may appear.

We fit the models by using the M-regression. The function \( \rho(u) \) used is the bi-square. Since no evidence of bad influential points is found, we check each model beginning with the slope test, which is a robust version equivalent to the F-test in OLS. The null and alternative hypotheses are \( H_0: \beta_1 = 0 \) and \( H_1: \beta_1 \neq 0 \).

For each model fitted in both galaxies, the robust slope test has a \( P \)-value less than 0.0001, so at 5% significance they are significant. The LMC V-band fitted model (26) has an \( R^2 \) value (the coefficient of determination) of 0.6248, meaning that the model explains about 62% of the variability of V. The LMC I-
The band model (27) and the LMC $W_f$ model (28) explain about 75% and 87% respectively of the observed variability.

We also use the robust Ramsey test (Ramsey 1969), given by the following equations:

$$y = \beta_1 + \beta_2 LP + \alpha_3 \hat{y}^2 + \varepsilon. \quad (29)$$

**Table 1**
Morphological Classification of OGLE-II Cepheids

|     | A1 | A2 | B1 | B2 | B3 | C1 | C2 | Total |
|-----|----|----|----|----|----|----|----|-------|
| LMC | 36 | 9  | 457| 127| 7  | 121| 8  | 765   |
| SMC | 10 | 13 | 458| 650| 42 | 128| 5  | 1306  |
Table 2
Bi-dimensional Classification of OGLE-II Cepheids in the Magellanic Clouds

| R.A.   | Decl. | ID        | Period | Class | R.A.   | Decl. | ID        | Period | Class |
|--------|-------|-----------|--------|-------|--------|-------|-----------|--------|-------|
| 05 33 02.35 | −70 15 33.3 | OGLE-LMC_SC1.25359 | 3.39729 | B1    | 05 27 57.84 | −69 53 48.5 | OGLE-LMC_SC3.44391 | 2.53947 | B1    |
| 05 27 43.07 | −70 00 49.7 | OGLE-LMC_SC3.26910 | 15.84364 | C1i   | 05 27 34.14 | −69 51 22.4 | OGLE-LMC_SC4.417847 | 7.49402 | C1i   |
| 05 24 57.16 | −69 13 32.8 | OGLE-LMC_SC5.455916 | 9.56961 | C1    | 05 19 38.13 | −69 37 44.6 | OGLE-LMC_SC7.388032 | 5.66859 | B1    |
| 05 15 37.55 | −69 30 29.4 | OGLE-LMC_SC8.112083 | 2.98282 | B2i   | 05 12 06.64 | −69 13 06.5 | OGLE-LMC_SC10.245266 | 3.96572 | B1    |
| 05 42 52.95 | −70 40 11.9 | OGLE-LMC_SC19.18756 | 3.48887 | B1    | 05 22 29.04 | −70 09 10.2 | OGLE-LMC_SC21.187856 | 3.16725 | B1i   |
| 00 36 55.83 | −73 56 27.2 | OGLE-SMC_SC1.1 | 14.3816 | C1i   | 00 42 34.89 | −73 25 37.2 | OGLE-SMC_SC3.12578 | 1.29508 | B2    |
| 00 44 51.77 | −73 33 56.5 | OGLE-SMC_SC3.178185 | 3.04334 | B2    | 00 45 42.73 | −72 52 39.7 | OGLE-SMC_SC4.38887 | 1.91467 | B1    |
| 00 54 32.93 | −72 30 26.5 | OGLE-SMC_SC6.324270 | 1.97277 | B2    | 00 55 18.59 | −72 43 12.2 | OGLE-SMC_SC7.120090 | 1.59513 | B2    |
| 01 01 49.84 | −72 29 55.5 | OGLE-SMC_SC9.73218 | 3.34991 | B1    | 01 05 03.87 | −72 51 52.3 | OGLE-SMC_SC10.78690 | 1.47799 | B2    |
| 01 08 17.81 | −72 44 17.8 | OGLE-SMC_SC11.75453 | 1.56745 | B1    | 01 09 04.87 | −72 20 14.5 | OGLE-SMC_SC11.117272 | 9.15889 | C1i   |

Note. The first two columns are the equatorial coordinates of the Cepheid (J2000.0 equinox). The third column indicates the OGLE-II ID. The fourth column gives the period in days. These data were reported for the LMC by Udalski et al. (1999a), and for the SMC by Udalski et al. (1999b). The last column gives the bi-dimensional class assigned in this study. The letter i identifies the influential Cepheids. A portion of the table is shown here for guidance regarding its form and content. The complete table is available in electronic form at the CDS’s VizieR service.

(This table is available in its entirety in machine-readable form.)
The null ($H_0$) and alternative ($H_1$) hypotheses are $H_0$: $\alpha_1 = 0$ and $H_1$: $\alpha_1 \neq 0$

$$y = \beta_1 + \beta_2LP + \alpha_1 y^2 + \alpha_2 y^3 + \varepsilon.$$  \hfill (30)

The null and alternative hypotheses are $H_0$: $\alpha_1 = \alpha_2 = 0$ and $H_1$: at least one $\alpha_i \neq 0$; $i = 1, 2$

$$y = \beta_1 + \beta_2LP + \alpha_1 y^2 + \alpha_2 y^3 + \alpha_3 y^4 + \varepsilon.$$  \hfill (31)

In Equations (29)–(31), $y$ is any of the magnitude vectors $V$, $I$, or the $W_I$ index. Moreover, values of $\alpha_i$ and $\beta_i$ are different for each photometric band.

This test shows that there is not enough evidence to reject $H_0$ due to its $P$-values at 5% significance. This implies that there is no evidence of specification problems in the LMC models. In contrast, for the SMC models there is enough evidence to reject $H_0$ due to $P$-values obtained by the test at the same level of significance. In particular, when Equation (29) is used, the LMC (SMC) $P$-values are 0.544 (0.011) for the $V$-band, 0.424 (0.006) for the $I$-band, and 0.434 (0.013) for the $W_I$ index. By using Equation (30), the LMC (SMC) $P$-values are 0.832 (0.015) for the $V$-band, 0.718 (0.006) for the $I$-band, and 0.714 (0.002) for the $W_I$ index. Finally, when Equation (31) is used, the LMC (SMC) $P$-values are 0.924 (0.013) for the $V$-band, 0.882 (0.005) for the $I$-band, and 0.709 (0.004) for the $W_I$ index.

### 4.1. Structural Breaks

In order to verify whether there are structural breaks (explained in Section 3.4) in model Equations (26)–(28), we fit the following model:

$$y = \beta_1 + \beta_2 \delta + \beta_3 LP + \beta_4 \delta LP + \varepsilon.$$  \hfill (32)

Each data set is split into two parts using a dummy variable represented by $\delta$, where

$$\delta = \begin{cases} 
1 & \text{if } P \leq 3.60 (2.57) \text{ day} \\
0 & \text{if } P > 3.60 (2.57) \text{ day}.
\end{cases}$$  \hfill (33)

The LMC (SMC) data are split around $P = 3.60 (2.57)$ days, in order to ensure that each data set has enough observations to avoid problems due to asymptotic results in robust regressions. As a result, the data sets have 240 (219) and 247 (235) points.

The hypotheses to contrast are $H_0$: $\beta_2, \beta_3 = 0$ and $H_1$: some $\beta_j \neq 0, j = 2, 4$. The parameters $\beta_2$ and $\beta_4$ are associated with $\delta$ in order to detect structural breaks in the zero point ($\beta_2$) and/or the slope ($\beta_4$) of the LL; if these parameters are not significant there is not enough evidence of structural breaks. In contrast, if one or both of them are significant there is enough evidence of structural breaks.

The obtained LMC $P$-values are 0.886 for the $V$-band, 0.775 for the $I$-band, and 0.572 for the $W_I$ index. Whereby, no evidence of structural breaks is found, hence the parameters obtained using $M$-estimators are stable at 5% significance.

We use the $M$-regression to estimate the LMC LL through the following equations:

$$V_0 = 17.005 - 2.683LP + e$$  \hfill (34)

$$I_0 = 16.549 - 2.930LP + e$$  \hfill (35)

$$W_I = 15.839 - 3.315LP + e.$$  \hfill (36)

Models explain about 62%, 75%, and 87% of the variability in the $V$-band, $I$-band and $W_I$, respectively. The confidence intervals at 95% for each $M$-estimator for the LMC zero point ($\beta_1$) and slope ($\beta_2$) are reported in Table 3. It is worth mentioning that the estimators for zero point and slope obtained by OLS are included between the confidence intervals obtained by $M$-regression.

To summarize, the robust tests have shown that there is no evidence to doubt the adequacy of the LMC models; as a result, they are acceptable.

For the SMC model the $P$-value obtained by the test of structural breaks is less than 0.001 for all bands, implying enough evidence to reject the null hypothesis and accept $H_1$ at 5% significance, i.e., there are structural breaks.

Because of the problems already detected in model Equations (26)–(28) for the SMC, further model analysis is needed using Equation (32). Therefore the individual significance of the parameters is analyzed. To do that, the following hypotheses are contrasted: $H_0$: $\beta_2 = 0$ and $H_1$: $\beta_2 \neq 0, j = 2, 4$. The individual robust test for $\beta_2$ shows it is significant at 5% significance, due to $P$-values of 0.035 in the $V$-band, 0.015 in the $I$-band, and 0.003 in the $W_I$ index. In contrast, the same test for $\beta_4$ is not significant at 5% significance since the associated $P$-values are 0.566 in the $V$-band, 0.381 in the $I$-band, and 0.090 in the $W_I$ index. There is not enough evidence to reject $H_0$, i.e., $\beta_4 = 0$. This means that the model has a different zero point for each data set but the same slope. Taking into account the specification problems and structural breaks detected in models (26)–(28), we fit the following model for SMC based on the previous individual significance analysis using (32).

$$y = \beta_1 + \beta_2 \delta + \beta_3 LP + \varepsilon.$$  \hfill (37)

Using OGLE-III Cepheids, Subramanian & Subramaniam (2015) determined the inclination of the SMC galaxy’s disk to be $64^\circ$. The geometrical distribution of the Cepheids across this disk could explain the behavior of the LL in the SMC, observed in Equation (37): linear regressions with the same slope and different zero points for each data set.

Model (37) is fitted by $M$-regression and no evidence of bad influential points is found. The robust Ramsey test is applied again to check whether there are specification problems in model (37). The equations and hypotheses of the test are

$$y = \beta_1 + \beta_2 \delta + \beta_3 LP + \alpha_1 y^2 + \varepsilon.$$  \hfill (38)

The null and alternative hypotheses are $H_0$: $\alpha_1 = 0$ and $H_1$: $\alpha_1 \neq 0$

$$y = \beta_1 + \beta_2 \delta + \beta_3 LP + \alpha_1 y^2 + \alpha_2 y^3 + \varepsilon.$$  \hfill (39)

The null and alternative hypotheses are $H_0$: $\alpha_1 = \alpha_2 = 0$ and $H_1$: at least one $\alpha_i \neq 0$; $i = 1, 2$

$$y = \beta_1 + \beta_2 \delta + \beta_3 LP + \alpha_1 y^2 + \alpha_2 y^3 + \alpha_3 y^4 + \varepsilon.$$  \hfill (40)

The test does not find enough evidence of specification problems because of the $P$-values obtained at 1% significance. In particular, for model (38) the SMC $P$-values are 0.933 for the $V$-band, 0.817 for the $I$-band, and 0.669 for the $W_I$ index. For model (39), the SMC $P$-values are 0.370 for the $V$-band, 0.224 for the $I$-band, and 0.040 for the $W_I$ index. For
Table 3
OGLE-II Leavitt Law for the LMC

| $V_0$ | $I_0$ | $W_I$ |
|-------|-------|-------|
| $17.005 \pm 0.064$ | $16.549 \pm 0.044$ | $15.839 \pm 0.029$ |
| $-2.683 \pm 0.108$ | $-2.930 \pm 0.074$ | $-3.315 \pm 0.049$ |
| $17.066 \pm 0.021$ | $16.593 \pm 0.014$ | $15.868 \pm 0.008$ |
| $-2.775 \pm 0.031$ | $-2.977 \pm 0.021$ | $-3.300 \pm 0.011$ |

Note. The reported $\beta$ estimators are those given by Equations (34)-(36) and were obtained by using the $M$-regression. Confidence intervals are reported at 95%. The last columns report the values of zero point and slope obtained by Udalski (2000) using OLS.

Table 4
OGLE-II Leavitt Law for the SMC

| $V_0$ | $I_0$ | $W_I$ |
|-------|-------|-------|
| $17.444 \pm 0.097$ | $17.019 \pm 0.074$ | $16.353 \pm 0.052$ |
| $0.135 \pm 0.069$ | $0.107 \pm 0.053$ | $0.063 \pm 0.037$ |
| $-2.551 \pm 0.147$ | $-2.869 \pm 0.110$ | $-3.342 \pm 0.079$ |
| $17.635 \pm 0.031$ | $17.149 \pm 0.025$ | $16.381 \pm 0.016$ |

Note. The reported $\beta$ estimators are those given by Equations (41)-(43) and were obtained by using the $M$-regression. Confidence intervals are reported at 95%. The last column reports the zero-point value obtained by Udalski (2000) using OLS.

Equation (40), the SMC $P$-values are 0.503 for the $V$-band, 0.346 for the $I$-band, and 0.091 for the $W_I$ index.

$M$-regression estimations for the SMC LL are shown in the following equations:

$$V_0 = 17.444 + 0.135\delta - 2.551LP + e$$  (41)
$$I_0 = 17.019 + 0.107\delta - 2.869LP + e$$  (42)
$$W_I = 16.353 + 0.063\delta - 3.34LP + e.$$  (43)

These models explain about 75%, 82%, and 87% of the variability in $V$, $I$, and the $W_I$ index, respectively. The confidence intervals at 95% for each $M$-estimator in Equations (41)-(43) are reported in Table 4. Besides, it is interesting that the confidence intervals for $M$-estimators include the values obtained by OLS, although there are slight differences.

To summarize, robust tests show specification problems and structural break points in the SMC models (26)-(28); therefore, analysis leads to model (37) as a better choice.

4.2. Universality of the LL

The universality hypothesis of the LL implies that the slope of the linear regression observed in the same filter in different galaxies has the same value, and shows a negligible dependence on the metallicity (García-Varela et al. 2013, and references therein). In order to test the universality of the LL, a linear model is proposed to fit simultaneously the data from the LMC and SMC. This model is given by

$$y = \beta_1 + \beta_2\delta + \beta_3LP + \beta_4\delta LP + \varepsilon$$  (44)

which is fitted in $V$, $I$, and the $W_I$ index after deleting the points previously discussed. The dummy variable $\delta$ indicates whether the observation is from the LMC or the SMC: $\delta = 1$ if a Cepheid is from the SMC and $\delta = 0$ otherwise. A total of 454 Cepheids from the SMC and 487 from the LMC are used to fit this model.

The model (44) is fitted by OLS. White’s test finds enough evidence to reject the null hypothesis of homoskedasticity, at 5% significance, since the $P$-value is 0.017 in $V$, less than 0.001 in $I$, and less than 0.0001 in the $W_I$ index. As happened before, OLS regression is inadequate due to heteroskedasticity problems and the presence of influential points. Besides, the distribution of residuals has a heavier tail than the normal distribution. Therefore, model (44) is fitted by $M$-regression using the bi-square function. No evidence of bad influential points is found; thus, the process to check the adequacy of the model continues.

The $P$-values of the slope test are less than 0.0001 in $V$, $I$, and for the $W_I$ index, so the models are significant at 5% significance. The hypotheses contrasted are $H_0$: $\beta_2 = \beta_3 = \beta_4 = 0$ and $H_1$: some $\beta_j \neq 0$, $j = 2, 3, 4$.

To test whether both galaxies have the same zero point and slope the following hypotheses are contrasted: $H_0$: $\beta_2 = \beta_4 = 0$ and $H_1$: some $\beta_j \neq 0$, $j = 2, 4$.

The $P$-values obtained by the robust test are less than 0.0001 in $V$, $I$, and for the $W_I$ index; therefore, there is enough evidence to reject the null hypothesis at 5% significance. Hence, the individual significance of each parameter must be analyzed. To do that, the following hypotheses are contrasted again: $H_0$: $\beta_2 = 0$ and $H_1$: some $\beta_j \neq 0$, $j = 2, 4$.

The robust test of individual significance for $\beta_2$ shows that it is significant at 5% significance, since its associated $P$-values are less than 0.0001 in $V$, $I$, and the $W_I$ index. The same test for $\beta_4$ is not significant at 5% significance because its associated $P$-values are 0.456 and 0.145 in $V$ and $I$, respectively. This means that the model has different zero points, as we expect because the galaxies are at different distances, but it has the same slope. As a result, the final equations fitted by the $M$-regression in the $V$- and $I$-bands, after deleting $\beta_4$, are

$$V_0 = 17.037 + 0.554\delta - 2.732LP + e$$  (45)
$$I_0 = 16.590 + 0.540\delta - 2.995LP + e.$$  (46)

Confidence intervals are reported in Table 5. Models (45) and (46) make an adjustment to the zero point when Cepheids are from the SMC but have the same slope for both galaxies at 5% significance. Moreover, they explain about 78% and 81% of the variability in $V$ and $I$, respectively.

The robust test of individual significance for $\beta_2$ in the $W_I$ index is significant at 5% significance because its associated $P$-value is 0.0014. This implies that each galaxy has its own regression line with different zero point and slope, and it explains about 87% of the $W_I$ variability. The following
equation gives the model fitted for the $W_I$ index:

$$W_I = 15.841 + 0.59463 - 3.318LP - 0.129dLP + e. \quad (47)$$

The confidence intervals at 95% for the estimated parameters in Equation (47) are $15.801 \leq \beta_1 \leq 15.880$ for zero point, $0.548 \leq \beta_2 \leq 0.639$ for zero-point adjustment when Cepheids are from the SMC, $-3.384 \leq \beta_3 \leq -3.251$ for slope, and $-0.209 \leq \beta_4 \leq -0.05$ for slope adjustment when Cepheids are from the SMC.

As a general result of this work, robust tests have shown that the LL of the LMC and SMC are universal in the $V$- and $I$-bands. There is enough evidence of two parallel regression lines, with the same slope at 5% significance but different zero points, as shown by models (45) and (46). However, for the $W_I$ index, it has been shown that there is no universal LL: the LMC and SMC have two completely different regression lines, as shown by Equation (47).

### 4.3. Analysis with OGLE-II and OGLE-IV Data

Based on the experience obtained by fitting the models (26) and (27), they are fitted for all $VI$ mean magnitudes of fundamental-mode Cepheids observed by the OGLE-II project. We also fit these models using the OGLE-IV fundamental-mode Cepheids, reported by Soszyński et al. (2015). It is worth mentioning that the published OGLE-IV mean magnitudes are not corrected for extinction. All the following models are fitted with the $MM$-regression because of its high BDP of 25%. The Least Trimmed Squares and Tukey function are used.

As no evidence is found of skewness in the models for both galaxies, we proceed with the analysis. Beginning with the LMC, model slope tests are significant, as can be seen in the $P$-values obtained: less than 0.0001 in the $V$- and $I$-bands (OGLE-II, OGLE-IV). Then, the test proposed by Godfrey & Orme (1994) is performed, giving $P$-values of 0.302 (OGLE-II) and 0.033 (OGLE-IV) in the $V$-band and 0.186 (OGLE-II) and 0.0003 (OGLE-IV) in the $I$-band; therefore, there is evidence of specification problems in the OGLE-IV I-band model at 1% significance. This could be due to blending problems and also to the lack of extinction correction in OGLE-IV data. To perform stability tests, the data are split by dummy variables into two data sets around $P = 3.85$ days (OGLE-II) and $P = 3.5877$ days (OGLE-IV), so as to have about half of observations in each set. No evidence of structural breaks is found, because the $P$-values obtained, by testing $\beta_2 = \beta_3 = 0$ versus the hypothesis that at least one of them is not 0 in Equation (32), are 0.601 (OGLE-II) and 0.453 (OGLE-IV) in the $V$-band, and 0.810 (OGLE-II) and 0.267 (OGLE-IV) in the $I$-band. Models explain about 61% and 68% of variability in the $V$- and $I$-bands respectively (OGLE-II), and 54% and 65% of variability in the $V$- and $I$-bands respectively (OGLE-IV). The estimated parameters of the models and their confidence intervals are reported in Table 6. Results for the $I$-band are shown only for information purposes because of their specification problems.

The OGLE-IV LMC data were also split into two around $P = 10$ days by Equation (32). No evidence of a structural break is found in the slope, because the $P$-value obtained is 0.705. However, the zero-point adjustment is significant. As a result, there is evidence of two parallel regression lines.

The SMC slope tests for models (26) and (27) are significant because the $P$-values obtained are less than 0.0001 for the $V$- and $I$-bands (OGLE-II, OGLE-IV). However, these models show problems. The test of Godfrey and Orme finds enough evidence of specification problems because the $P$-values obtained are less than 0.0001 for the $V$- and $I$-bands (OGLE-II, OGLE-IV). To perform stability tests, data are split around 2.0039 days (OGLE-II) and 1.9429 days (OGLE-IV) so as to have about half of observations in each data set. Enough evidence of structural breaks is found: $P$-values are less than 0.0001 in both bands (OGLE-II) when testing $\beta_2 = \beta_3 = 0$ versus the hypothesis that at least one of them is not 0 in Equation (32). Besides, when individual significance tests are performed for $\beta_2$ and $\beta_4$, the $P$-values obtained are less than 0.0001 for $\beta_2$ in both bands, and 0.0322 and 0.0082 for $\beta_4$ in the $V$-band and $I$-band respectively. As a result, in both bands there are two regression lines at 5% significance (OGLE-II).

Unlike OGLE-II models for the SMC, there is not enough evidence to support the idea that OGLE-IV model parameters for the slope are unstable with $P$-values 0.784 and 0.597 in the $V$- and $I$-bands respectively, while the zero-point adjustment, $\beta_2$ (Equation (32)), is still significant with a $P$-value less than 0.0001 in both bands. As a result, stability tests show that there are two parallel regression lines in both bands. New models are fitted based on the individual significance analysis using model (32) for the $V$- and $I$-bands. However, the test of Godfrey and Orme shows that specification problems persist. Nevertheless, for information purposes only, we report the estimated values of the slope at 95% confidence intervals for the OGLE-IV data: $-2.780 \pm 0.055$ for the $V$-band and $-3.024 \pm 0.046$ for the $I$-band.

Summarizing, the LMC models fitted with OGLE-II and OGLE-IV data are adequate except for the $I$-band model (OGLE-IV). The adequacy of the OGLE-II models could be explained because only 36% of Cepheids in the LMC have intermediate and large dispersion. Therefore, as Cepheids with small dispersion are the largest fraction, the exclusion of those with intermediate and large dispersion has no significant impact on the estimated slope value, because confidence interval for slope ($\beta_2$), shown in Table 6, includes the slope estimation ($\beta_2$) shown in Table 3. Moreover, LMC models fitted with all OGLE-IV data have confidence intervals for slope that contain estimations obtained by using the models fitted with all OGLE-II data (see $\beta_2$ values in Table 6).

On the other hand, there is enough evidence to show that SMC models are inadequate. Unlike the LMC, the largest fraction of OGLE-II SMC data (~64%) are for Cepheids with intermediate and large dispersion; as a result, when models are fitted with all data, those Cepheids form the majority of the sample. The presence of a large fraction of Cepheids with intermediate and large dispersion in the SMC galaxy can be explained by the geometry of this galaxy, as discussed in Section 4.1. It is interesting that SMC models improve when they are fitted with OGLE-IV data because, despite

### Table 5

| $V_I$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|-------|-----------|-----------|-----------|
| 17.037 ± 0.047 | 0.554 ± 0.030 | -2.732 ± 0.074 |
| 16.590 ± 0.034 | 0.540 ± 0.022 | -2.995 ± 0.054 |

Note. The reported $\beta$ estimators are those given by Equations (45) and (46), which were obtained by using the $MM$-regression. Confidence intervals are reported at 95%.
specification problems, there is not enough evidence of instability of the slope parameter in both bands.

5. CONCLUSIONS

Our main conclusions can be summarized as follows.

By applying the traditional OLS and robust statistical linear regression models to optical OGLE-II and OGLE-IV data, it is possible to affirm that the problems of break and nonlinearity of the LL disappear for the LMC when M- or MM-regressions are used. In that case, the LL is obtained without breaks in slope, as shown in Equations (45) and (46), and it is not necessary to exclude or reject points from the data sample. The models in this case are adequate and do not present specification problems or structural breaks, except in the OGLE-IV I-band, probably due to the lack of extinction correction among other causes. These facts allow us to conclude that the nonlinearity of the LMC LL is a consequence of using an inadequate statistical method.

For the SMC, despite the use of robust regressions, problems in specification, structural breaks, and adequacy are found. This could be caused in part by blending problems due to the geometry of this galaxy, which could be generating a large fraction of Cepheids with intermediate and large dispersion, as discussed in Section 4.1. This astronomical reason seems to imply that Cepheid variables in galaxies with elongated structures like that of the SMC are not appropriate to fit a reliable LL. In particular, for the SMC, we found that excluding the Cepheids with intermediate and large dispersion, i.e., stars that exhibit the communalities of influential Cepheids, avoids the problems of adequacy and specification. This could be indicating that this kind of Cepheid really has influence in the fitting of the LL for galaxies with similar geometry to that of the SMC. However, from a statistical point of view, there is no reason to exclude these stars from the analysis.

The OLS method has violations of its assumptions for the LMC and SMC galaxies, with OGLE-II and OGLE-IV data, as discussed at the beginning of Section 3.1. In particular, two important problems are worthy of mention here. The first one is the fact that rejecting outliers using the sigma-clipping algorithm is not adequate because the precision of the parameters is adjusted artificially, and besides there are violations of OLS assumptions. Without strong non-statistical reasons to support the rejection of points, this method is invalid in the context of LL linear regression. The second problem is the presence of influential points, which implies the need to find communalities in them. In this work we found that a fraction of Cepheids with a grade of moderate or large dispersion are influential points. The communalities exhibited by them are suggested by the bumps in their light curves, i.e.,

Cepheids showing the Hertzsprung progression, or eclipsing variations, or evidence of changes in period. These influential Cepheids were excluded from the sample using OLS regression in order to solve adequacy problems; however, new influential points appear that again imply violations of the OLS assumptions.

Although the obtained M-estimators (slope and zero point) and their respective confidence intervals include the OLS estimators reported by Udalski (2000), the relevance of using the M-estimators is that they allow us to make a reliable statistical inference, since none of their assumptions is violated, unlike the OLS estimations.

When Cepheids from the LMC and SMC are combined to fit a single model, in each of the V- and I-bands, the linear relations remain valid after omitting the Cepheids already discussed. The M-model shows that each galaxy has its own regression line and they differ only in the zero point. This implies that the OGLE-II LL is universal in the range of metallicities of the Magellanic Clouds. In contrast, the combined model for Wj shows that each galaxy has its own linear regression with the expected difference in zero point, but with an unexpected difference in slope. This result has already been reported in the literature, but its causes are not yet clear.

LMC models fitted by using all OGLE-II and OGLE-IV data by MM-regression are adequate, but the OGLE-IV I-band has specification problems. These problems could be due to blending and the lack of extinction correction in these data. There is no evidence of instability in their zero-point and slope parameters. On the other hand, despite the specification problems in SMC models for OGLE-II and OGLE-IV data, there is not enough evidence of instability in the slope parameters when they are fitted with OGLE-IV data, which have more observations than OGLE-II. However, when OGLE-IV data are split at P = 10 days, there is evidence of two parallel regression lines.

It is clear from our results that influential Cepheids do not have an important effect on the LL for the LMC. However, they could have an impact in the case of galaxies with similar geometry to that of the SMC. The effect of these Cepheids in the measurement of distances to these galaxies will be studied in a forthcoming paper.

Finally, based on the results of this work, we suggest the use of the LL relations given by Equations (45) and (46) as a universal law for the SMC and LMC galaxies. We also suggest using the slope and zero point for the V-band, obtained for the LMC using OGLE-IV data, as shown in Table 6. For the SMC using OGLE-II without the influential Cepheids, we suggest using Equations (41) and (42).

| OGLE-II | OGLE-IV | Soszyński et al. |
|--------|--------|-----------------|
| V      | 17.029 ± 0.041 | 17.436 ± 0.025 | 17.438 ± 0.012 |
|        | −2.732 ± 0.063 | −7.011 ± 0.028 | −2.690 ± 0.018 |
| I      | 16.553 ± 0.029 | 16.823 ± 0.018 | 16.822 ± 0.009 |
|        | −2.947 ± 0.043 | −2.922 ± 0.029 | −2.911 ± 0.014 |

Note. The reported β estimators are those given by Equations (26) and (27). They were obtained by using the MM-regression. Confidence intervals are reported at 95%. The last columns report the values of zero point and the slope obtained by Soszyński et al. (2015) using the sigma-clipping algorithm and OLS over OGLE-IV data. The OGLE-II and OGLE-IV MM-estimators obtained in the V-band and their respective confidence intervals include at 1.0σ the OLS estimators reported by Udalski (2000) and Soszyński et al. (2015).
The authors thank the referee, Dr. Pawel Pietrukowicz, for his important recommendations and suggestions that have improved this manuscript. A.G.V., B.E.S., and J.R.M. acknowledge support from Facultad de Ciencias, Universidad de los Andes, through Proyecto Semilla. S.V.D. acknowledges the Physics Department, Universidad de los Andes, for the fellowship supporting the postdoctoral position held within the Astronomy group.

REFERENCES

Aikawa, T., & Whitney, C. 1984, ApJ, 282, 527
Aikawa, T., & Whitney, C. 1985, ApJ, 296, 165
Alcock, C., Allsman, R., Alves, D., et al. 2002, ApJ, 573, 338
Babcock, H. 1958, ApJS, 3, 141
Bono, G., Marconi, M., & Stellingwerf, R. 2000, A&A, 360, 245
Butkovskaya, V., Plachinda, S., Baklanova, D., & Butkovskyi, V. 2014, IAUS, 301, 393
Caputo, F., Marconi, M., & Musella, M. 2000, A&A, 354, 610
Christy, R. 1968, QJRAS, 9, 13
Drapper, N., & Smith, H. 1998, Applied Regression Analysis (3rd ed.; New York: Wiley)
Fadeyev, Y. 2013, AstL, 39, 746
Feigelson, E. D., & Babu, G. J. 2012, Modern Statistical Methods for Astronomy with R Applications (1st ed.; New York: Cambridge Univ. Press)
Fernie, J. D. 1969, PASP, 81, 707
Fouqué, P., Ariarragada, P., Storm, J., et al. 2007, A&A, 476, 73F
Freedman, W., & Madore, B. 2010, ARA&A, 48, 673
García-Varela, A., Sabogal, B., & Ramírez-Tannus, M. 2013, MNRAS, 431, 2278
Gastine, T., & Dintrans, B. 2008, A&A, 490, 743
Gieren, W., Pilecki, B., Pietrzyński, G., et al. 2014, ApJ, 786, 80
Gieren, W., Pilecki, B., Pietrzyński, G., et al. 2015, ApJ, 815, 28G
Gieren, W., Storm, J., Barnes, T., et al. 2005, ApJ, 627, 224G
Godfrey, L. G., & Orme, C. D. 1994, International Economic Review, 48, 489
Harris, J., & Zaritsky, D. 2006, AJ, 131, 2514
Hertzsprung, E. 1926, BAN, 3, 115
Kafadar, K. 1983, JRNBS, 88, 105
Kukarkin, B. 1937, Astr. Zhurnal U.S.S.R., 14, 125
Kukarkin, B. 1937, Astr. Zhurnal U.S.S.R., 14, 125
Leavitt, H. S., & Pickering, E. C. 1912, HarCi, 173, 1
Madore, B., & Freedman, W. 1991, PASP, 103, 933
Molnár, L., & Szabados, L. 2014, MNRAS, 442, 3222
Montgomery, D. C., Peck, E. E., & Vining, G. G. 2012, Introduction to Linear Regression Analysis (5th ed.; New York: Wiley)
Neilson, H., & Ignace, R. 2014, A&A, 563, L4
Ngeow, C.-C., Kanbur, S. M., & Bellinger, E. P., et al. 2012, Ap&SS, 341, 105
Nun, I., Pichara, K., Protopapas, P., et al. 2014, ApJ, 793, 23
Pietrzyński, G., Gieren, W., Udalski, A., et al. 2004, A&J, 128, 2815
Pietrzyński, G., Thompson, I., Gieren, W., et al. 2010, Natur, 468, 542
Pietrzyński, G., Thompson, I., Graczyk, D., et al. 2011, ApJ, 742, 20
Pilecki, B., Graczyk, D., & Gieren, W. 2015, ApJ, 809, 29
Plachinda, S. 1990, Krymskaia Astrofiyizicheskaia Observatoriia, 81, 112
Plachinda, S. 1990, Krymskaia Astrofiyizicheskaia Observatoriia, 81, 112
Plachinda, S. 2000, A&A, 360, 642
Poleski, R. 2008, AcA, 58, 313
Ramsey, J. B. 1969, Journal of the Royal Statistical Society Series B (Methodological), 31, 350
Ramsey, J. B. 1974, in Frontiers in Econometrics, ed. P. Zarembla (New York: Academic Press), 13
Sandage, A., & Tammann, G. 1968, ApJ, 151, 531
Sandage, A., & Tammann, G. 1969, ApJ, 157, 683
Schwarzenberg-Czerny, A. 1989, MNRAS, 241, 153
Severny, A., Kuvshinov, V., & Nikulin, N. 1974, Krymskaia Astrofiyizicheskaia Observatoriia, Izvestiia Obs, 50, 3
Simon, N., & Schmidt, E. 1976, ApJ, 205, 162
Soszyński, I., Poleski, R., Udalski, A., et al. 2008, A&A, 58, 163
Soszyński, I., Udalski, A., Poleski, R., et al. 2012, A&A, 62, 219
Soszyński, I., Udalski, A., Szymański, M., et al. 2015, A&A, 65, 279
Stothers, R. 2009, ApJL, 696, L37
Strassmeier, K. 2009, A&ARv, 17, 251
Subramanian, S., & Subramaniam, A. 2015, A&A, 573, A135
Tammann, G. A., Sandage, A., & Reindl, B. 2003, A&A, 404, 423
Turner, D., Abdel-Sabour Abdel-Latif, M., & Berdnikov, L. 2006, PASP, 118, 410
Udalski, A. 2000, A&A, 50, 279
Udalski, A., Soszyński, I., Szymański, M., et al. 1999a, A&A, 49, 223
Udalski, A., Soszyński, I., Szymański, M., et al. 1999b, A&A, 49, 437
Udalski, A., Szymański, M., Kamiński, M., et al. 1999c, A&A, 49, 201
Udalski, A., Szymański, M., & Szymański, G. 2015, A&A, 65, 1
Udalski, A., Wyrzykowski, Ł., Pietrzyński, G., et al. 2001, A&A, 51, 221
Wade, G., Chadid, M., Shorlin, S., et al. 2002, A&A, 392, L17
Weiss, W. 1986, A&A, 160, 243
White, H. 1980, Econometrica, 48, 817
Whitney, C. 1983, ApJ, 274, 830
Yohai, V. 1987, AnSta, 15, 642

The Astrophysical Journal, 824:74 (14pp), 2016 June 20
GARCÍA-VARELA ET AL.