Cereal Biotechnology and Biochemistry

Online ISSN: 2783-5170

Homepage: https://cbb.razi.ac.ir

# Linear modeling regression analysis: A mini-review over regression pros and cons

**Armin Saed-Moucheshi[1] ✉, Soodabeh Saedi[2], Fatemeh Ansarshourijeh[3], Abbas Rezaizad[1] & Amin Sadeghi[4]**

[1] Crop and Horticulture Research Department, Kermanshah Agricultural and Natural Resources Research and Education Center (AREEO), Kermanshah, Iran.

[2] Department of Plant Protection, College of Agriculture, Razi University, Kermanshah, Iran.

[3] Department of Natural Resources Engineering, College of Agriculture, Shiraz University, Shiraz, Iran.

[4] Department of Plant Protection, College of Agriculture, University of Kurdistan, Sanandaj, Iran.

✉ Corresponding author. E-mail: saedmoocheshi@gmail.com

## ABSTRACT

**Introduction:** Powerful and practical statistical packages have simplified the analysis and thus developed the application of data science in all research fields. Accordingly, regression has been applied to almost all aspects of the life sciences. However, misuse of this model has been reported in the past decades. This article aims to examine modeling with this important statistical method and introduce readers to the correct use of this method.

**Materials and methods:** This review article uses real data, and the supplementary materials provide the method for performing the regression analysis in SAS and R statistical software and their related codes.

**Results:** In the required assumptions of the regression model, the residuals of the model must be normally distributed, but performing the normality test for the actual values of the response variable or any of the explanatory variables is not mandatory. Therefore, researchers should not obsess more than necessary about the normal distribution of real data. On the other hand, almost all normality test methods, such as Kolmogorov-Smirnov, are designed for large numbers of data, typically more than a thousand samples. This suggests that using such methods to test the normality of model residuals estimated from a small number of data, mostly less than a hundred cases, would be inaccurate. Another issue regarding applying the regression model is related to the co-linearity of the explanatory variables. There are still signs of correlation in a data set where all variables are generated separately and randomly in a statistical package. This means that it is very hard to find a correlation coefficient equal to zero (r = 0) even between any pair of separate, random variables. Therefore, in all regression models, there are some kinds of correlation between explanatory variables, but the important issue here is that only high correlation causes severe problems in the model. For collinearity test it would be better to use specialized methods such as Variance Inflation Factor (VIF) or Principal Component Analysis (PCA). The linearity of the model is one other assumption of regression model. Data transformation might be helpful under the situation of non-linearity of the model. However, transformation changes the variables unit, altering the array direction in a geometric space. Researchers should be careful regarding the use of modeling a large number of data affects the probability values in variance analysis due to increasing the value of the degree of freedom of the model.

**Conclusion:** As the number of data points increases, the degree of freedom of the error term increases rapidly. Therefore, the final error mean squared significantly reduces. In contrast, the scatter of data points around the regression line may be too wide. For this reason, using the coefficient of determination, usually called (R-Squared), is a suitable criterion for testing the model's fit. High coefficient values indicate a suitable model for the data set used. It should be noted that in a multiple regression model, the higher the number of explanatory variables used in the model, the higher the value of this coefficient increases. For such conditions, when the number of explanatory variables is large, another form of this coefficient, called the adjusted coefficient of determination (adjusted $R^2$), has been introduced. The use of this coefficient in the approximations creates a limit on the number of variables used in the regression model. Accordingly, the number of variables in the model as explanatory variables should not exceed the number of samples (or the number of tens) in a set, and researchers should avoid using more variables than the number of samples.

**Keywords:** Multiple Regression, Durbin-Watson, Error Mean Squares, Model Residuals, Residual Normal Distribution.

**Article Type:** Review Article

**Article history: Received:** 21 Feb 2024, **Revised:** 18 Apr 2024, **Accepted:** 23 May 2024, **Published online:** 21 Jun 2024

**Cite this article:** Saed-Moucheshi, A., Saedi, S., Ansarshourijeh, F., Rezaizad, A. & Sadeghi, A. (2024). Linear modeling regression analysis: A mini-review over regression pros and cons. *Cereal Biotechnology and Biochemistry*, *3*(2), 346-360. DOI: 10.22126/cbb.2024.11198.1086



10.22126/cbb.2024.11198.1086

**Publisher:** Razi University

بیوتکنولوژی و بیوشیمی غلات

شاپا الکترونیکی: ۵۱۷۰-۲۷۸۳

Homepage: https://cbb.razi.ac.ir

# مدل‌سازی خطی و تحلیل رگرسیون: مروری کوتاه بر مزایا و معایب مدل‌سازی خطی و شرایط استفاده از مدل رگرسیون

آرمین ساعدموچشی[۱]✉، سودابه ساعدی[۲]، فاطمه انصارشوریجه[۳]، عباس رضایزاد[۱] و امین صادقی[۴]

[۱] بخش تحقیقات علوم زراعی و باغی، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی کرمانشاه، مرکز تحقیقات، آموزش و ترویج کشاورزی، کرمانشاه، ایران.

[۲] بخش گیاه‌پزشکی، دانشکده کشاورزی، دانشگاه رازی، کرمانشاه، ایران.

[۳] گروه مهندسی منابع طبیعی، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران.

[۴] گروه گیاه‌پزشکی، دانشکده کشاورزی، دانشگاه کردستان، سنندج، ایران.

✉نویسنده مسئول. رایانامه: saedmoocheshi@gmail.com

## چکیده

**مقدمه:** امروزه بسته‌های نرم‌افزاری قدرتمند و کاربردی، تحلیل داده‌ها را ساده کرده و در نتیجه کاربرد علم داده را در تمام زمینه‌های تحقیقاتی توسعه داده است. بر این اساس، رگرسیون تقریباً در تمام جنبه‌های علوم‌زیستی، از سلامت انسان گرفته تا کشاورزی و علوم دامی اعمال شده است. اما در دهه‌های گذشته اشتباهات شایان توجهی در استفاده از این مدل گزارش شده است. هدف از این مقاله بررسی مدل‌سازی با این روش مهم آماری و آشنا کردن خوانندگان جهت کاربرد درست این روش و مفروضات و شرایط استفاده از آن است.

**مواد و روش‌ها:** در این مقاله مروری از داده‌های واقعی استفاده گردیده است و نحوه انجام تحلیل‌های انجام شده در نرم‌افزارهای آماری SAS و R و کدهای مربوط به آنها در قسمت پیوست آورده شده است.

**یافته‌ها:** در مفروضات مورد نیاز مدل رگرسیونی، باقیمانده‌های مدل باید به طور نرمال توزیع شده باشند، اما انجام آزمون نرمال بودن برای مقادیر واقعی متغیر پاسخ یا هر یک از متغیرهای مستقل اجباری نیست. از سوی دیگر، تقریباً تمام روش‌های تست توزیع نرمال، مانند Kolmogorov-Smirnov، برای تعداد زیاد داده، طراحی شده‌اند. این نشان می‌دهد که استفاده از چنین روش‌هایی برای آزمون نرمال بودن باقیمانده‌های مدل تخمین زده شده بر اساس تعداد داده پایین، عمدتاً کمتر از صد مورد، چندان دقیق نخواهد بود. موضوع دیگر مربوط به هم‌خطی بین متغیرهای مستقل است. باید به این نکته توجه کرد که یافتن ضریب همبستگی برابر با صفر ( R = 0) حتی بین هر جفت متغیر تصادفی جداگانه بسیار دشوار است. بنابراین در تمامی مدل‌های رگرسیونی به نوعی همبستگی بین متغیرهای مستقل وجود خواهد داشت، اما موضوع مهم این است که فقط همبستگی زیاد باعث ایجاد مشکلات شدید در مدل می‌شود. پیشنهاد می‌گردد که به جای استفاده از روش ساده همبستگی از روش‌های تخصصی مانند ضریب تورم واریانس (VIF) یا تجزیه و تحلیل مؤلفه اصلی (PCA) برای تشخیص شدت هم‌خطی استفاده گردد. یکی دیگر از مفروضات رگرسیون مربوط به خطی بودن مدل است که گاهی تبدیل این مشکل را برطرف کند. باید توجه شود که تبدیل داده‌ها منجر به تغییر واحد متغیرها یا تغییر جهت برداری آنها در یک فضای هندسی و در برخی موارد تغییر ساختار صحیح آنها می شود.

**نتیجه‌گیری:** در مدل رگرسیون با افزایش تعداد داده، درجه آزادی خطا به سرعت افزایش می یابد و میانگین مجذور خطای نهایی به میزان قابل توجهی کاهش می یابد. مقدار کم میانگین‌مربعات‌خطا منجر به یک مدل بسیار معنی‌دار می‌شود. در مقابل، پراکندگی نقاط داده در اطراف خط رگرسیون ممکن است بسیار گسترده باشد. به همین دلیل، استفاده از ضریب‌تبیین که معمولاً معیار مناسبی برای تست برازش مدل است. هرچه پراکندگی نقاط مربوط به داده‌ها در اطراف خط رگرسیون گسترده‌تر باشد، مقدار ضریب تعیین کمتر است. مقادیر بالای این ضریب نشان دهنده مدل مناسب برای مجموعه داده‌های مورد استفاده است. یک مقدار مناسب برای ضریب‌تبیین را نمی توان بین دامنه‌ای از مقادیر برای همه آزمایش‌ها توصیه کرد.

**واژه‌های کلیدی:** آزمون دوربین واتسون، باقیمانده‌های مدل، توزیع نرمال باقیمانده، رگرسیون چندگانه، میانگین مربعات خطا.

**نوع مقاله:** مقاله مروری

**نوع مقاله: دریافت:** ۱۴۰۳/۰۱/۰۸ **اصلاح:** ۱۴۰۳/۰۱/۲۱ **پذیرش:** ۱۴۰۳/۰۲/۱۳، **انتشار آنلاین:** ۱۴۰۳/۰۴/۰۱

**استناد**: ساعدموچشی، آ.، ساعدی، س.، انصارشوریجه، ف.، رضایزاد، ع. و صادقی، ا. (۱۴۰۳). مدل‌سازی خطی و تحلیل رگرسیون: مروری کوتاه بر مزایا و معایب مدل‌سازی خطی و شرایط استفاده از مدل رگرسیون. *بیوتکنولوژی و بیوشیمی غلات*، ۳(۲)، ۳۴۶-۳۶۰. DOI: 10.22126/cbb.2024.11198.1086



ناشر: دانشگاه رازی

**مقدمه**

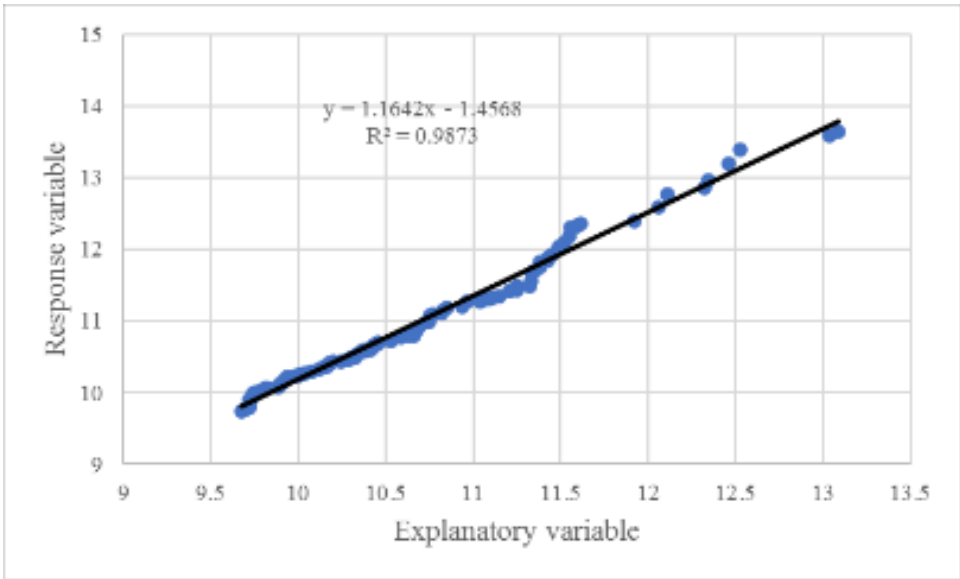
در طول قرن گذشته پیشرفت‌های زیادی در استخراج داده‌ها از منابع بیولوژیکی و غیربیولوژیکی صورت گرفته است. افزایش تعداد داده‌ها منجر به معرفی تکنیک های جدید و توسعه تکنیک‌های قدیمی برای تجزیه و تحلیل آن‌ها و استخراج رابطه ساختاری بین ویژگی‌های و متغیرهای مختلف در یک مجموعه داده بر اساس قوانین ریاضی شده است. جهت شناسایی درست رابطه بین ویژگی‌های مختلف بر اساس قواعد ریاضی، باید نوع و شدت این روابط را درک کرد. پس از شناسایی روابط و ساختار، مدل‌سازی‌ها و تخمین‌های مختلف می‌توانند برای یافتن بهترین برازش در مجموعه داده اعمال گردیده و آزمایش شوند. هنگامی که روابط به مدل تبدیل شدند و ضرایب مربوطه بر این اساس ساخته شد، می‌توان از آن‌ها برای پیش بینی در موقعیت‌های مختلف استفاده کرد (Saed-Moucheshi *et al.*, 2013b). یکی از کاربردی‌ترین و محبوب‌ترین روش‌ها برای شناسایی روابط بین متغیرها و مدل‌سازی این رابطه‌ها، رگرسیون خطی است که می‌تواند به درستی برای پیش‌بینی مجموعه داده‌های مورد نظر یا داده‌های جدید استفاده شود. در این مقاله سعی شده است مبانی مدل‌سازی خطی رگرسیون و کاربرد آن در تمایز ساختار مجموعه داده به طور واضح توضیح داده شده و تمامی مشکلات و استفاده های نادرست که تاکنون اعمال شده است مورد بحث قرار گیرد.

**۲. مدل سازی رابطه بین متغیرهای پاسخ و متغیرهای توضیحی**

در علم داده روش‌های مختلفی برای مدل‌سازی و یافتن رابطه بین متغیرهای مختلف وجود دارد. محققین انواع مختلفی از آزمایش‌ها و مشاهدات را برای بررسی روابط سببی (علت و معلول) بین متغیرها و یافتن مهم‌ترین ویژگی‌هایی که ممکن است به طور مؤثر به آن‌ها در ایجاد تغییرات در متغیر پاسخ[1] کمک کند، اعمال کرده‌اند. از نظر مدل‌سازی و رابطه در آمار، دو عبارت متغیر پاسخ و مستقل را می‌توان معرفی نمود. متغیر مستقل ویژگی است که محقق برای ارزیابی تأثیر آن بر متغیر پاسخ مورد ارزیابی قرار می‌دهد. به عنوان مثال، دوزهای مختلف دارو یا سمی که برای ارزیابی دوز دقیق کشندگی اعمال می‌شود متغیر توضیحی است در حالی که متغیر پاسخ ویژگی است که سعی دارد بر اساس تغییرات در متغیر توضیحی، مدل یا پیش بینی شود و در این مثال مقدار کشندگی دارو به عنوان متغیر پاسخ مطرح است. به عنوان یک مثال دیگر، کشندگی یک دارو و سطح بیان یک ژن می‌تواند به عنوان متغیر پاسخ در ارتباط با (در پاسخ به) محرک های دیگر مانند تغیرات هورمونی و شرایط محیطی به عنوان متغیرهای توضیحی مطرح باشد (Saed-Moucheshi *et al.*, 2013a).

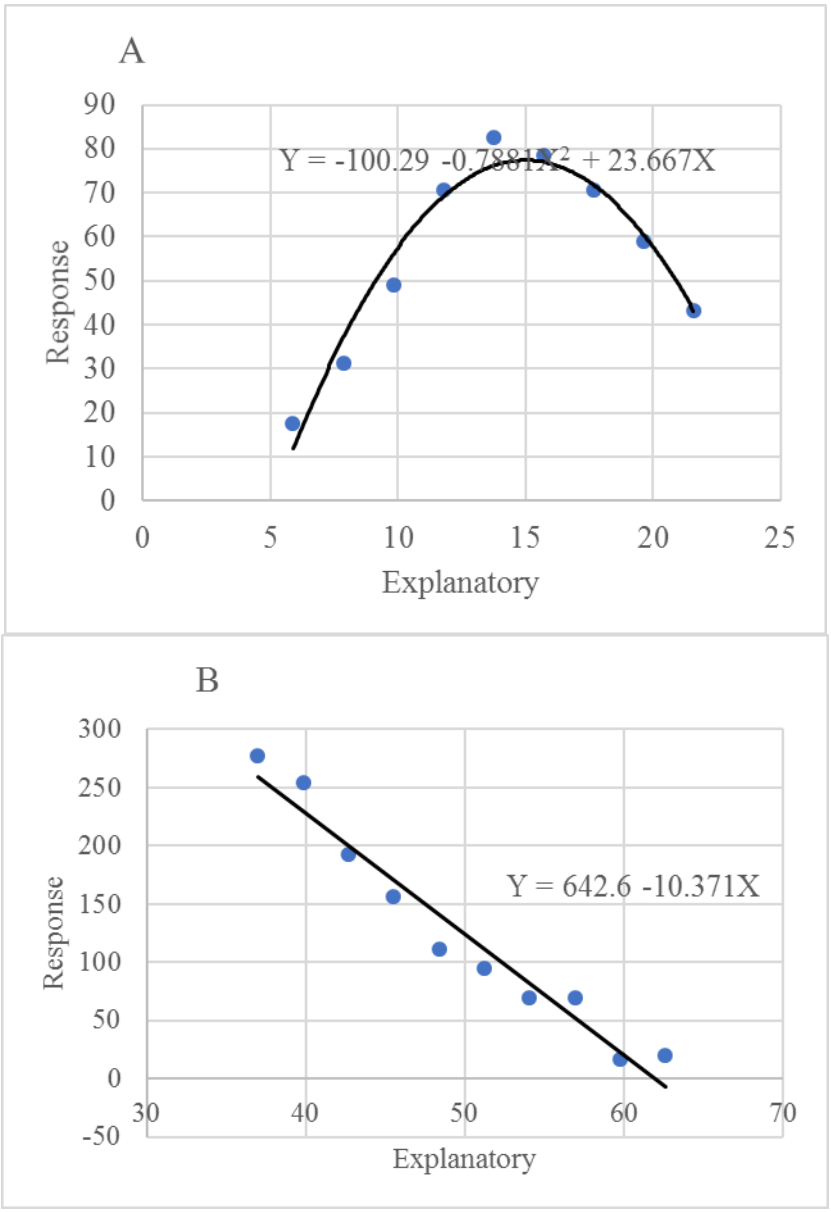
[1] Response Variable

رگرسیون خطی شامل یک متغیر پاسخ است که تحت تأثیر سایر متغیرهای مورد بررسی قرار دارد و در آن یک متغیر توضیحی (رگرسیون ساده) یا چندگانه (رگرسیون چندگانه) بر متغیر پاسخ تأثیر می‌گذارد (Aliakbari *et al.*, 2013). گاهی اوقات، متغیر پاسخ را متغیر وابسته و متغیر توضیحی را متغیر مستقل نیز می‌نامند. در همین حال، متغیر پاسخ ممکن است تحت تأثیر متغیرهایی غیر از متغیرهایی باشد که در مدل رگرسیونی در نظر گرفته می‌شوند، همچنین ممکن است در مدل رگرسیونی یک یا چند متغیر توضیحی تأثیر معناداری در مدل نداشته باشند. بنابراین، واژه متغیرهای پاسخ و توضیحی برای استفاده به جای اصطلاحات وابسته یا مستقل مناسب‌تر است. اصطلاح خطی برای نشان دادن خطی بودن مدل رابطه به کار می‌رود. مدل خطی و رابطه خطی گاهی با یکدیگر متفاوت هستند. رابطه خطی برای زمانی اعمال می‌شود که رابطه بین متغیرهای پاسخ و توضیح مشابه یک خط مستقیم باشد (Saed-Moucheshi *et al.*, 2019) (شکل ۱). از سوی دیگر، در یک مدل خطی، رابطه بین متغیرهای پاسخ و توضیحی ممکن است خطی یا غیر خطی باشد، در حالی که خود مدل خطی است (معادله ۱ و معادله ۲). در یک مدل خطی، اجزای مدل یک علامت ریاضی منفی یا مثبت به مؤلفه‌های موجود قبلی اضافه می‌کنند، در حالی که هیچ ضرب یا تقسیمی بین مؤلفه‌ها وجود ندارد (James *et al.*, 2021). شکل ۲ یک مدل خطی (شکل ۲الف) و یک رابطه غیرخطی (شکل ۲ب) بین متغیرهای پاسخ و توضیحی را نشان می‌دهد.




**شکل ۱. نمودار پراکندگی متغیر پاسخ در مقابل متغیر توضیحی بر اساس مدل رگرسیون خطی ساده.**

**Figure 1. scatterplot of the response vs. explanatory variables based on a simple linear regression model.**




**شکل ۲. نمودار پراکندگی متغیر پاسخ در مقابل متغیر مستقل در مدل خطی با رابطه غیر خطی (B) و مدل خطی با رابطه خطی (B)**

**Figure 2. Scatter plot of response vs. explanatory for linear model with non-linear relationship (A) and Linear model with linear relationship (B)**

رگرسیون خطی ساده (معادله ۱):

$$Y = a + bX + e \quad (۱)$$

$$Y = \alpha + \beta X + \varepsilon \quad (۲)$$

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (۳)$$

(۱) مدل مربوط به یک نمونه

(۲) مدل مربوط به یک جمعیت حقیقی

(۳) این مدل جهت بسط دادن مدل به تعداد متغیرات بیشتر مورد استفاده قرار گرفته است

رگرسیون خطی چندگانه (معادله ۲):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_i X_i + \ldots + \varepsilon$$

Y = متغیری که می خواهید پیش بینی کنید (متغیر وابسته یا پاسخ).

X = متغیری که برای پیش بینی استفاده می‌شود (متغیر مستقل)

a، α، و β0 = عرض از مبدا

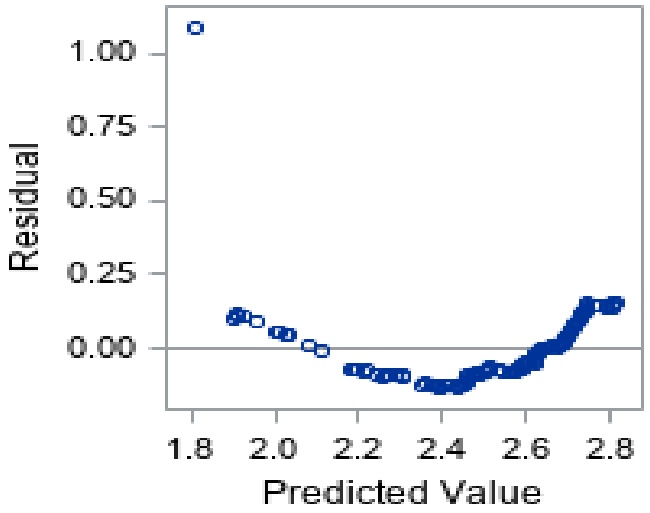
b و βi = شیب خطوط رگرسیون

e و ε = باقیمانده رگرسیون (خطای مدل)

## ۳. مدل رگرسیون و مفروضات آن

### ۱.۳. فرض خطی بودن مدل

مدل رگرسیون خطی به دلیل مفروضاتی که باید در مجموعه داده‌های مورد استفاده و ویژگی‌های آن‌ها وجود داشته باشد، همیشه قابل اجرا نیست. همانطور که قبلا ذکر شد، مدل رابطه بین متغیرهای پاسخ و توضیح باید افزایشی (یا خطی) باشد. این یک فرض مهم است زیرا در شرایط غیرخطی بودن مدل، نتایج پیش‌بینی‌شده در مدل به احتمال زیاد نادرست بوده و منجر به نتیجه‌گیری اشتباه می‌شود (Lio & Liu, 2018). پرکاربردترین روش برای ارزیابی خطی‌بودن یک مدل، ترسیم مقادیر باقیمانده در مقابل مقادیر برازش شده (مقادیر پیش‌بینی‌شده متغیر پاسخ) است (شکل ۳).




**شکل ۳. مقدار باقیمانده در مقابل مقادیر پیش بینی شده بر اساس مدل رگرسیون خطی نامناسب**

**Figure 3. Residual value vs. predicted values based on an unappropriated linear regression model**

در شرایطی که رابطه یا مدل خطی نیست، متخصصین علم داده مجاز به انجام تبدیل و نرمال سازی داده‌های هستند (Vosough *et al.*, 2015). انواع مختلفی از روش‌های تبدیل داده‌ها و نرمال سازی وجود دارد، مانند لگاریتم گرفتن، معکوس کردن داده‌ها، به توان رساندن، مجموعه تبدیل‌های Box-Cox و غیره. تبدیل Box-Cox یک روش تبدیل داده مؤثر است و به شدت در نمونه‌های دنیای واقعی کاربرد دارد زیرا تقریباً همه روش‌های دیگر تبدیل داده را در بر می‌گیرد (Astivia & Zumbo, 2019). این تبدیل داده مجموعه‌ای از تبدیل‌های توانی (به توان رساندن داده‌ها) به عنوان مثال ۳- تا ۳+ بر اساس درجه یا گام است. بر این اساس داده‌های مورد بررسی در یک گام به عنوان مثال ۰/۱ از توان ۳- تا توان ۳+ به ترتیب تبدیل می‌شوند و مورد بررسی قرار می‌گیرند. رساندن داده‌ها به توان ۱ منجر به ایجاد داده‌های اصلی می‌شود، در حالی که اگر داده‌ها به توان ۱- برسند مقادیر داده‌ها را به مقادیر معکوس تبدیل می‌کند. به توان رساندن داده به توان عدد ۰/۵ باعث رادیکال‌گیری یا مجذورگیری از داده‌ها می‌شود. پس از انجام تبدیل داده، تجزیه و تحلیل‌ها بر روی مجموعه جدید داده-های تبدیل شده انجام می‌شود (Souza *et al.*, 2017).

### ۲.۳. استقلال باقی‌مانده‌های مدل

شرط دیگر برای مجاز بودن در استفاده از مدل رگرسیون، عدم وجود همبستگی سریالی و متقابل بین داده‌ها است که به وسیله بررسی خود هم‌بستگی بین خطاهای (باقی‌مانده-های) مدل مشخص می‌شود. خود هم بستگی[1] زمانی اتفاق می‌افتد که بتوان مقادیر باقی‌مانده از یک داده (خطای مربوط به داده دوم) را به وسیله باقی‌مانده از داده قبلی (خطای داده اول) تخمین زد. خود همبستگی بین مقادیر خطا به شدت دقت مدل نهایی را کاهش می‌دهد. در برخی موارد، استفاده از روش‌های تبدیل داده ممکن است مشکل خود همبستگی داده‌ها را حل کند (Bazilevsky, 2018). آماره دوربین-واتسون[2]، که گاهی اوقات به صورت DW (برای حروف اول نویسندگان) به آن اشاره می‌شود، معمولاً برای اندازه گیری مقدار خود همبستگی استفاده می‌شود. اکثر نرم افزارها می‌توانند به راحتی این محاسبات را انجام دهند و بر اساس این آماره، مقادیر DW بین اعداد صفرتا چهارقرار دارد (0<DW<4). هر چه مقدار DW به ۲ نزدیک‌تر باشد، خودهمبستگی ضعیف‌تر خواهد بود. محدوده $0 < DW < 2$ (DW بین صفر و دو) وجود یک خودهمبستگی مثبت را نشان می دهد و محدوده $2 < DW < 4$ نیز یک خود همبستگی منفی را نشان می دهد. برخی از محققان ممکن است ۱/۵ <DW< ۲/۵ را به عنوان

[1] Autocorrelation
[2] Durbin-Watson statistics

مناسب ترین محدوده DW انتخاب کنند و برخی ممکن است محدوده یک تا سه را مناسب بدانند، با این حال، بهترین محدوده بهتر است به عدد دو نزدیکتر باشد (Kabaila *et al.*, 2021).

### ۳.۳. داده پرت و واریانس خطا

با ارائه یک مدل رگرسیون برای متغیر پاسخ بر اساس متغیرهای توضیحی، می‌توان مقادیر پاسخ را تخمین زد که به آنها مقادیر پیش بینی نیز گفته می‌شود. تفاوت بین مقادیر واقعی متغیرهای پاسخ و مقادیر پیش‌بینی شده آنها نشان دهنده مقادیر خطا یا باقیمانده است (جدول ۱). هرچه عبارت خطا کوچکتر باشد، مدل دقیق تر است. با مربع کردن (به توان ۲ رساندن) مقادیر خطا (باقیمانده) و سپس جمع همه مقادیر خطا با هم که بعد از به توان رسیدن به اعداد مثبت تبدیل شده‌اند می‌توان واریانس خطا یا میانگین مربعات خطاه[1] (MSE) را برای مدل رگرسیون تخمین زد. از این آماره می‌توان به عنوان معیاری برای تخمین مقدار دقت مدل استفاده کرد. همچنین، جهت بررسی اثر متغیرهای توضیحی و به اصطلاح بررسی سطح معنی داری (احتمال[2]) آنها نیز از این آماره استفاده می‌گردد و ضریب برآوردی به وسیله مدل مورد آزمون قرار می‌گیرند. در آمار، فرض بر این است که هر مقدار پیش‌بینی‌شده از متغیر پاسخ دارای یک مقدار منحصربه‌فرد است و این مقدار نماینده میانگین یک جمعیت متشکل از تمامی مقادیر متغیر پاسخی است که دارای یک مقدار ثابت از متغیرهای توضیحی هستند. بر این اساس، هر مقدار خطای تخمینی نیز نشان دهنده جمعیتی از مقادیر خطا (یا میانگین خطا) در نقطه در نظر گرفته شده برای متغیرهای توضیحی است. بنابراین، یک واریانس خطای جداگانه برای برای هر یک از مقادیر پیش‌بینی‌شده بر اساس مدل رگرسیون می‌تواند از نظر تئوری جهت آزمایش مقدار برازش مدل استفاده گردد. اصطلاح ناهمسانی واریانس یا هتروسکداستیکی[3] نشان دهنده واریانس غیر ثابت در عبارات خطای برآورد شده توسط مدل رگرسیون است. به عبارت دیگر، واریانس تخمین زده شده در هر نقطه پاسخ باید برابر یا تقریباً برابر با سایر واریانس‌های برآورد شده برای سایر نقاط پاسخ باشد. وجود نقاط پرت (نقاط داده با تفاوت یا فاصله زیاد از سایر نقاط مجموعه داده) معمولاً منجر به واریانس غیر ثابت و ایجاد مشکل ناهمسانی می شود (Baum & Lewbel, 2019). لازم به ذکر است که مقادیر پرت[4] تأثیر زیادی بر عملکرد مدل نشان می‌دهند و باید در هر برآورد از مدل رگرسیونی ارزیابی شوند. استفاده از نمودار پراکندگی مقادیر باقیمانده در مقابل مقادیر برازش شده (مقادیر پیش‌بینی شده توسط مدل برای متغیر پاسخ) روشی مناسب برای یافتن واریانس غیر ثابت یا ناهمسانی واریانس خطا است. بر اساس شکل ۴ الف، اگر نقاط در نمودار پراکندگی بین دو خط موازی قرار

[1] Error Mean Squares
[2] p-value
[3] Heteroscedasticity
[4] Outliers

گیرند، ناهمسانی وجود نخواهد داشت. هر شکل دیگری برای نقطه باقیمانده در چنین نمودار پراکندگی نشان دهنده وجود ناهمسانی در مدل است (شکل ۴).

### ۴.۳. نرمالیتی باقی‌مانده‌ها

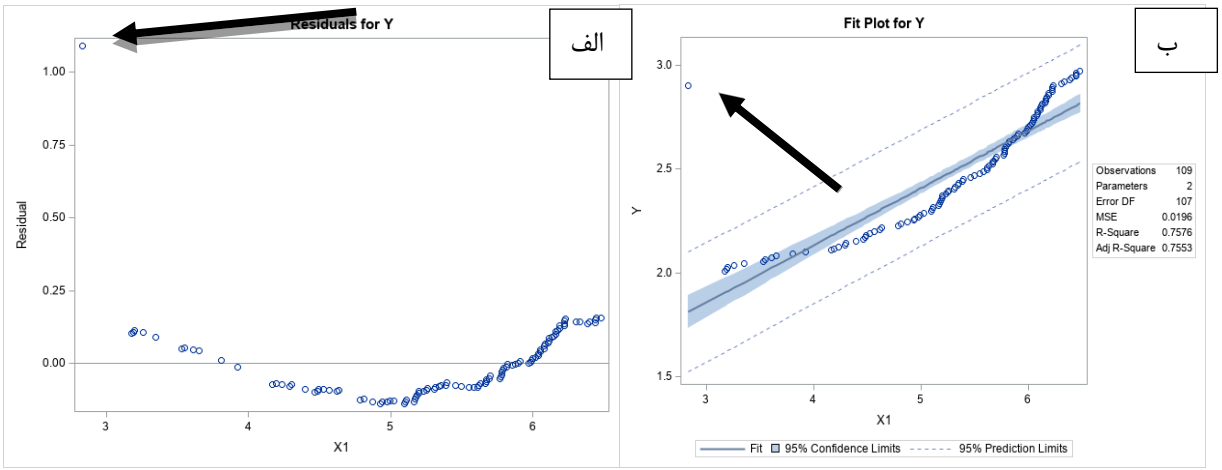
توزیع نرمال عبارات خطا[1] یکی دیگر از مفروضات مهم مدل رگرسیون است. همانطور که توزیع غیر نرمال عبارات خطا رخ می دهد، فواصل اطمینان ضرایب تخمین زده شده در مدل به طور غیرعادی کم‌تر یا گسترده تر از ضرایب واقعی می‌شود. گاهی اوقات، وجود توزیع غیرعادی برای مقادیر خطای مدل رگرسیون، وجود چند نقطه داده غیرعادی را نشان می‌دهد که باید دقیقاً ارزیابی شوند. با این حال، در برخی موارد، توزیع واقعی عبارات خطا در واقع به طور معمول توزیع نمی شود و رگرسیون غیر خطی بهتر است برازش شود. انجام تست نرمال بودن بر روی مقادیر خطا مانند روش کولموگروف-اسمیرنوف در کنار رسم باقیمانده های مدل در برابر باقی‌مانده‌های قابل تخمین با استفاده از نمودار چارکی یا QQ پلات[2] به شناسایی توزیع عبارات خطا کمک می‌کند (شکل ۵).

[1] Normality of Residuals
[2] Quantile-Quantile Plot

**جدول ۱. نمایش مقادیر باقیمانده (خطا) داده بر اساس مدل رگرسیون.**

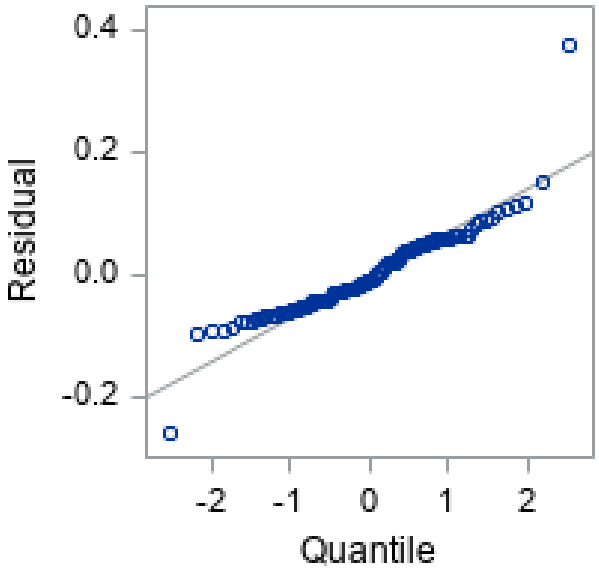
**Table 1. Representing the residual (error) values of the data point based on the regression model.**

| مشاهده<br>Observation | مقدار پاسخ واقعی<br>Real response values | مقدار پاسخ تخمینی<br>Estimated response values | مقدار خطا (باقیمانده)<br>Error values (residuals) |
|---|---|---|---|
| 1 | 2.9 | 2.5234 | 0.3766 |
| 2 | 2.01 | 2.2706 | -0.2606 |
| 3 | 2.02 | 2.0608 | -0.0408 |
| 4 | 2.03 | 2.0165 | 0.0135 |
| 5 | 2.04 | 2.0779 | -0.0379 |
| 6 | 2.05 | 1.8941 | 0.1559 |
| 7 | 2.05 | 2.0021 | 0.0479 |
| 8 | 2.06 | 2.0607 | -0.0007 |
| 9 | 2.07 | 2.1489 | -0.0789 |
| 10 | 2.08 | 2.1726 | -0.0926 |
| 11 | 2.09 | 2.1712 | -0.0812 |
| 12 | 2.1 | 2.1685 | -0.0685 |
| 13 | 2.11 | 2.1145 | -0.0045 |
| 14 | 2.12 | 2.1008 | 0.0192 |
| 15 | 2.13 | 2.1583 | -0.0283 |
| 16 | 2.14 | 2.119 | 0.021 |
| 17 | 2.14 | 2.0511 | 0.0889 |
| 18 | 2.15 | 2.148 | 0.002 |
| 19 | 2.16 | 2.1026 | 0.0574 |
| 20 | 2.17 | 2.14 | 0.03 |




**شکل ۴. تشخیص نقاط پرت با استفاده از نمودار پراکندگی پاسخ در مقابل متغیر توضیحی (الف) و باقیمانده در مقابل متغیر توضیحی (ب).**

**Figure 4. Detecting outliers by using scatter plot of response vs. explanatory (A) and residuals vs. explanatory (B) variables.**




**شکل ۵. نمودار چندکی(Quantile-quantile plot) (QQ-plot) باقیمانده بر اساس مدل رگرسیون برای تشخیص نوع توزیع باقیمانده(residuals' distribution).**

**Figure 5. Quantile-quantile plot (QQ-plot) of residuals based on the regression model for detecting the type of residuals' distribution.**

### ۵.۳. استقلال متغیرهای توضیحی

علاوه بر خودهمبستگی بین عبارات خطا در مدل های رگرسیونی، یک اصطلاح همبستگی دیگر به نام هم-خطی[10] چندگانه نیز ممکن است وجود داشته باشد. هم-خطی زمانی اتفاق می‌افتد که همبستگی بالایی بین متغیرهای توضیحی وجود داشته باشد (Morrissey & Ruxton, 2018). وقوع هم‌خطی منجر به ایجاد خطای بیشتر در مدل رگرسیون و همچنین دشواری در تخمین سهم واقعی هر متغیر توضیحی در پاسخ را به دنبال دارد. برای رهایی از این مشکل، دانشمندان علم‌داده ممکن است جدول همبستگی بین متغیرهای توضیحی را در نظر بگیرند و آنهایی را که مقادیر همبستگی بالایی دارند حذف کنند. در برخی موارد، زمانی که وابستگی‌های زیادی بین متغیرهای توضیحی وجود دارد، استفاده از رگرسیون مؤلفه‌اصلی[11] (PCR به عنوان حروف اول) ممکن است مشکل حل کند (Saed-Moucheshi *et al.*, 2013a). در PCR، متغیر پاسخ براساس اجزائی که از تجزیه به مؤلفه‌های اصلی[12] به دست آمده است مدل-سازی گردد و از خود متغیرهای توضیحی به صورت مستقیم استفاده نگردد. همچنین روش‌های دیگری برای مقابله با مشکل چندهم خطی مانند استفاده از رگرسیون ریج[13] (RR) و رگرسیون وزنی وجود دارد که از نظر اصولی مشابه هستند. نمودار پراکندگی ساده بین متغیرهای توضیحی به سادگی می‌تواند مقدار همبستگی بین این متغیرها را نشان داده و به تعیین هم‌خطی کمک کنند (Bagya Lakshmi *et al.*, 2018). ضریب تورم واریانس[14] یا VIF روش دیگری برای آزمایش هم‌خطی بین متغیرهای توضیحی است. بر این اساس، مقدار VIF $\leq 4$ (کوچکتر یا مساوی ۴) نشان دهنده نبود هم‌خطی چندگانه است در حالی که یک مقدار VIF بالا هم‌خطی قوی بین متغیرها را پیشنهاد می کند (بیشتر VIF $\geq 10$).

[10] Collinearity
[11] Principal Component Regression
[12] Principal Component Analysis
[13] Ridge Regression
[14] Variance Inflation Factor

## ۴. نتیجه‌گیری و نکات قابل توجه جهت استفاده از مدل رگرسیون

بسته‌های آماری قدرتمند و کاربردی، تحلیل‌ها را ساده کرده و در نتیجه کاربرد علم داده را در تمامی زمینه‌های تحقیقاتی توسعه داده است. بر این اساس، رگرسیون تقریباً در تمام جنبه‌های علوم زیستی از سلامت انسان گرفته تا کشاورزی و علوم دامی اعمال شده است. با این حال، استفاده نادرست از این مدل در دهه های گذشته گزارش شده است. همانطور که در مفروضات مورد نیاز مدل رگرسیون ذکر شد، باقیمانده های مدل باید به طور نرمال توزیع شوند ولی انجام آزمون نرمالیتی برای مقادیر واقعی متغیر پاسخ یا هر یک از متغیرهای توضیحی اجباری نیست. بنابراین، محققان نباید بیش از حد مورد نیاز نسبت به توزیع نرمال داده‌های واقعی وسواس داشته باشند. از سوی دیگر، تقریباً تمام روش‌های تست نرمال،

مانند Kolmogorov-Smirnov، برای تعداد زیادی داده، طراحی شده‌اند. این موضوع نشان می‌دهد که استفاده از چنین روش‌هایی برای آزمایش نرمال بودن باقیمانده‌های مدل برآورد شده از تعداد کمی داده، عمدتاً کمتر از صد مورد، چندان دقیق نخواهد بود. بنابراین، در مدل‌سازی مجموعه داده‌های کوچک و تحلیل رگرسیون، احتمال آزمون نرمالیتی ممکن است کمتر از ۰/۰۵ شود. در حالی که توزیع واقعی آنها ممکن است نرمال باشد، علاوه بر این، یک آزمایش، داده‌های جمع‌آوری شده از یک جامعه واقعی نیست، بلکه نمونه ای است که برای نشان دادن خصوصیات جامعه جمع‌آوری شده است. این بدان معناست که اگر توزیع جمعیتی که نمونه‌برداری شده یا در مدل رگرسیونی مورد مطالعه قرار می گیرد، مطمئناً نرمال باشد، انجام تست های مختلف نرمال بودن بر روی باقیمانده های نمونه اجباری نیست، زیرا داده‌ها از جمعیتی با توزیع نرمال می‌آیند. موضوع دیگر در مورد کاربرد مدل رگرسیون مربوط به هم‌خطی متغیرهای توضیحی است. در مجموعه داده‌ای که همه متغیرهای آن به طور جداگانه و به صورت تصادفی در یک بسته آماری تولید می شوند، هنوز نشانه هایی از همبستگی وجود دارد. این بدان معنی است که پیدا کردن ضریب همبستگی برابر با صفر ($R = 0$) حتی بین هر جفت متغیرهای جداگانه و تصادفی بسیار سخت است. بنابراین در همه مدل‌های رگرسیونی باید نوعی همبستگی بین متغیرهای توضیحی وجود داشته باشد، اما مسئله مهم در اینجا است که فقط همبستگی زیاد باعث ایجاد مشکل شدید در مدل می‌شود. بر این اساس، استفاده صرف از جدول همبستگی صرفاً برای دلالت بر هم‌خطی چندگانه راه مناسبی نیست، در عوض، پژوهشگران بهتر است از روش‌های تخصصی مانند ضریب تورم واریانس (VIF) یا تحلیل مؤلفه‌های اصلی (PCA) برای تشخیص شدت هم-خطی استفاده کنند. از طرفی، برخی از مجموعه داده‌ها قادر به برآوردن مفروضات مربوط به مدل‌سازی خطی نیستند. در چنین شرایطی، استفاده از تبدیل داده ممکن است مشکل را حل کند. اگرچه استفاده از هر نوع تبدیل برای برخی از متغیرها در نوع خاصی از مجموعه داده‌ها مجاز نیست. تبدیل داده‌ها منجر به تغییر واحد متغیرها یا تغیر جهت بردار آنها در یک فضای هندسی و در برخی موارد تغییر ساختار صحیح آنها می شود. علاوه بر این، مجموعه‌های از داده‌ها وجود دارند که برای تحلیل رگرسیون مناسب نیستند، حتی اگر همه انواع تکنیک‌های تبدیل برای آنها اعمال شود. به عنوان مثال، داده‌های دودویی[15] که داده‌های متشکل از مقادیر صفر و یکهستند، برای مدل‌سازی رگرسیون خطی قابل استفاده نیستند و باید با استفاده از نوع خاصی از تحلیل که رگرسیون لجستیک[16] نامیده می‌شود، مدل‌سازی شوند. بنابراین، محققان باید در استفاده و مدیریت تبدیل داده‌های خود مراقب باشند. در برخی موارد، استفاده از مدل‌سازی تعداد زیاد داده مقادیر احتمال در تحلیل واریانس را به دلیل بالا بردن مقدار درجه آزادی مدل مورد تاثیر قرار می‌دهد. متعاقباً، ایت تاثیر می‌تواند روی معنی‌داری هر یک از

[15] binary
[16] Logistic Regression

ضرایب مدل تاثیر گذار باشد و در نهایت مشکلاتی را در مدل‌سازی و برآورد برای محققین به وجود آورد. این مشکل در واقع ریشه در روش محاسبه واریانس خطا و همچنین درجه آزادی آن است. با افزایش تعداد نقاط داده، درجه آزادی مربعات میانگین خطا به سرعت افزایش می‌یابد. بنابراین، میانگین مربعات نهایی خطا به طور چشمگیری کاهش می‌یابد. مقدار کم میانگین مربعات خطا منجر به یک مدل بسیار معنادار می شود. در مقابل، پراکندگی نقاط داده در اطراف خط رگرسیون ممکن است بسیار گسترده باشد. به همین دلیل استفاده از ضریب‌تبیین[۱۷] معیار مناسبی برای آزمون تناسب مدل است. هرچه نقطه داده در اطراف خط رگرسیون گسترده تر شود، مقدار ضریب تعیین کمتر می شود. مقادیر بالای این ضریب نشان دهنده یک مدل مناسب برای مجموعه داده استفاده شده است. یک مقدار مناسب برای ضریب‌تبیین را نمی‌توان بین دامنه‌ای از مقادیر برای همه آزمایش‌ها توصیه کرد، با این حال در بیشتر موارد، مقدار بالاتر از ۰/۷ در آزمایشات علوم کشاورزی نشان دهنده تناسب مدل به داده‌ها است. مقدار ضریب تعیین کمتر ۰/۴ در آزمایشات کشاورزی در بیشتر موارد نشان دهنده یک مدل نامناسب است. علاوه بر این، باید توجه داشت که در یک مدل رگرسیون چندگانه، هر چه تعداد متغیرهای مستقل استفاده شده در مدل بیشتر باشد، مقدار این ضریب به میزان بیشتری افزایش می‌یابد. برای چنین شرایطی، زمانی که تعداد متغیرهای مستقل زیاد باشد، شکل دیگری از این ضریب که به آن ضریب‌تبیین تعدیل شده[۱۸] نامیده می شود، معرفی شده است. استفاده از این ضریب باعث در تقریب‌ها باعث ایجاد محدودیتی در مورد تعداد متغیرهای استفاده شده در مدل رگرسیون ایجاد می‌کند. بر این اساس، تعداد متغیرهای موجود در مدل به‌عنوان متغیرهای مستقل نباید از تعداد نمونه‌ها در یک مجموعه تجاوز کند و محققان باید از استفاده از تعداد متغیرهای بیشتر از تعداد نمونه‌ها خودداری کنند.

[17] Coefficient of Determination (R-Squared)

[18] Adjusted R-Square

**References**

Aliakbari, M., A. Saed-Moucheshi, H. Hasheminasab, H. Pirasteh-Anosheh, M. T. Asad and Y. Emam. 2013. Suitable stress indices for screening resistant wheat genotypes under water deficit conditions. International journal of Agronomy Plant Production, 4(10), 2672-2695

Astivia, O. L. O. and B. D. Zumbo. 2019. Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS. Practical Assessment, Research & Evaluation, 24**,** 265-279.

Bagya Lakshmi, H., M. Gallo and R. M. Srinivasan. 2018. Comparison of regression models under multi-collinearity. Electronic Journal of Applied Statistical Analysis, 11(1)**,** 340-368.

Baum, C. F. and A. Lewbel. 2019. Advice on using heteroskedasticity-based identification. The Stata Journal, 19(4)**,** 757-767.

Bazilevsky, M. P. 2018. Research of new criteria for detecting first-order residuals autocorrelation in regression models. Mathematics and Mathematical Modeling, 6(3)**,** 13-25.

James, G., D. Witten, T. Hastie, R. Tibshirani, G. James, D. Witten, T. Hastie and R. Tibshirani. 2021. Linear model selection and regularization. An introduction to statistical learning: with applications in R, 10, 225-288.

Kabaila, P., D. Farchione, S. Alhelli and N. Bragg. 2021. The effect of a Durbin–Watson pretest on confidence intervals in regression. Statistica Neerlandica, 75**,** 4-23.

Lio, W. and B. Liu. 2018. Residual and confidence interval for uncertain regression model with imprecise observations. Journal of Intelligent & Fuzzy Systems, 35(2)**,** 2573-2583.

Morrissey, M. B. and G. D. Ruxton. 2018. Multiple regression is not multiple regressions: the meaning of multiple regression and the non-problem of collinearity. Philosophy, Theory, and Practice in Biology, 10(3)**,** 563-588.

Saed-Moucheshi, A., E. Fasihfar, H. Hasheminasab, A. Rahmani and A. Ahmadi. 2013a. A review on applied multivariate statistical techniques in agriculture and plant science. International journal of Agronomy and Plant Production, 4, 127-141.

Saed-Moucheshi, A., M .Pessarakli and B. Heidari. 2013b. Comparing relationships among yield and its related traits in mycorrhizal and nonmycorrhizal inoculated wheat cultivars under different water regimes using multivariate statistics. International Journal of Agronomy, 13(13), 345-365.

Saed-Moucheshi, A., H. Razi, A. Dadkhodaie, M. Ghodsi and M. Dastfal. 2019. Association of biochemical traits with grain yield in triticale genotypes under normal irrigation and drought stress conditions. Australian Journal of Crop Science, 13(2), 272-295.

Souza, L. C., R. M. C. R. Souza, G. J. A. Amaral and T. M. Silva Filho. 2017. A parametrized approach for linear regression of interval data. Knowledge-Based Systems, 131, 149-159.

Vosough, A., R. Ghouchani and A. Saed-Moucheshi. 2015. Genotypic Variation and Heritability of Antioxidant related Traits in Wheat Landraces of Iran. Biological Forum, 7(2)**,** 43-55