Journal of Statistical Sciences, Spring and Summer, 2023
Vol. 17, No. 1, pp 81-102
DOI: 10.52547/jss.17.1.5

Analysis of High Dimensional Data Using Development Support Vector Regression, Functional Regression, Ridge and Lasso Regression

Rouhi, A., Jahadi, F, Roozbeh, M., Zalzadeh, S.,
Department of Statistics, Semnan University, Semnan, Iran.

Corresponding author: M. Roozbeh, mahdi.roozbeh@semnan.ac.ir

Received: 20/4/2022 **Revised:** 27/1/2023 **Accepted and Published Online:** 30/1/2023.

Introduction

Regression models have attracted particular attention in econometrics, engineering, psychology, biology and other areas. Nowadays, many real-world data sets carry structures in which the number of covariates may greatly exceed the sample size, called high-dimensional problems. In such situations, several researches have been pursued addressing forecasting a response variable, estimating an underlying vector parameter and selecting the variables. This study tried to model these data sets by introducing practical approaches such as Support Vector Regression (SVR), functional regression, ridge, and lasso regression methods. It is tried to apply a regression model with penalized principal components on a high-dimensional data set. Then, the generalized SVR is used on the transformed functional data set. SVR is a way to fit a regression model, which is an incredible member of the machine learning family. SVR has been established to be an efficient technique in real-value function estimation. As a supervised-learning approach, SVR trains using a symmetrical loss function, penalizing high and low misestimates equally. To evaluate the effectiveness of the proposed methods in practice, some numerical experiments are made on riboflavin production and simulated data sets to shed light on the practical performance of the suggested method.

Material and Methods

SVR uses the same principles as the support vector machine for classification, with only a few minor differences. Nowadays, due to the extension

of data types and modernization in data storage, the functional data set is very observable. To analyze these types of data sets, the discrete data set must be converted to the continuous data set using the smoothing approach. Then, the principal component method is applied to the developed curves by a smoothing technique to decrease the number of features. The principal component technique is one of the data reduction approaches with the idea of reducing the dimensions and conserving as much information as possible from the explanatory variables. Also, the data sets are analyzed by SVR (based on linear, polynomial, sigmoid and radial kernels) and generalized SVR extended by cross-validation. Then, we compare the fitting results using correlation squared, mean squared error and mean absolute error percentage deviation criteria.

Results and Discussion
Exploring high-dimensional data sets is difficult because classical methods cannot be used to estimate and interpret them. Therefore, we have to use alternative methods to analyze them. SVR, functional regression, LASSO and ridge regression are some of the best ways to be applied in such cases. In this study, based on the numerical result, generalized SVR and then SVR with linear kernel were better than LASSO and ridge estimation.

Conclusion
Because of the ill conditionality of $X^\top X$ matrix، the high-dimensional data analysis was not possible with classical methods. So، in this study، we tried to apply alternative approaches such as SVR، functional regression، LASSO and ridge regression. The generalized SVR method with linear kernel was the best method to model and predict the high-dimensional riboflavin data set. In the simulated data set، the generalized SVR method with radial kernel was reasonably efficient in contrast to the other methods.

Keywords: Functional regression, High dimensional data, Lasso regression, Ridge regression, Support vector regression.

Mathematics Subject Classification (2010): 62G08, 62H25.

مجله علوم آماری، بهار و تابستان ۱۴۰۲
جلد ۱۷، شماره ۱، ص ۸۱ – ۱۰۲
DOI: 10.52547/jss.17.1.5
مقاله پژوهشی

تحلیل داده‌های با بعد بالا با استفاده از رگرسیون بردار پشتیبان تعمیم یافته، رگرسیون تابعی، رگرسیون ستیغی و لاسو

آرتا روحی، فاطمه جهادی، مهدی روزبه و سعید زال‌زاده
گروه آمار، دانشکده آمار، دانشگاه سمنان.

نویسنده مسئول : مهدی روزبه، mahdi.roozbeh@semnan.ac.ir
تاریخ دریافت: ۱۴۰۱/۱/۳۱ تاریخ بازنگری: ۱۴۰۱/۱۱/۷ تاریخ پذیرش و انتشار: ۱۴۰۱/۱۱/۱۰

چکیده: تحلیل داده‌های با بعد بالا با استفاده از روش‌های رگرسیون کلاسیک انجام پذیر نیست و ممکن است نتایج آن گمراه کننده باشد. در این تحقیق سعی شده است با معرفی تکنیک‌های جدید و قدرتمندی مانند رگرسیون بردار پشتیبان، رگرسیون تابعی، رگرسیون ستیغی و لاسو، به واکاوی این‌گونه داده‌ها پرداخته شود. در این راستا، با تحلیل دو مجموعه داده بعد بالا (داده‌های مربوط به تولید ریبوفلاوین و شبیه‌سازی شده) با روش‌های معرفی شده، به ارزیابی کاراترین مدل با استفاده از سه معیار (مجذور همبستگی، میانگین توان دوم خطا و میانگین انحراف درصد خطای مطلق) با توجه به نوع داده‌ها پرداخته می‌شود.
واژه‌های کلیدی: داده‌های با بعدبالا، رگرسیون بردار پشتیبان، رگرسیون تابعی، رگرسیون ستیغی، رگرسیون لاسو.

نویسنده مسئول : مهدی روزبه، mahdi.roozbeh@semnan.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08 ، 62H25.

۱ مقدمه

امروزه جمع آوری داده‌ها با شیوه‌های متفاوتی انجام می‌شود. گاهی اوقات با توجه به نوع داده، کاهش هزینه در جمع آوری داده و مواردی از این قبیل داده‌های با بعد بالا به وجود می‌آیند. بدین معنا که تعداد متغیرهای تبیینی(p)

از تعداد مشاهدات (n) بیشتر است(افرون و هستی ، ۲۰۱۷). انجام رگرسیون خطی بر روی اینگونه داده‌ها با مشکلاتی همراه و نتایج حاصل از آن گمراه کننده است، زیرا برای برآورد ضرایب به شیوه کمترین توان دوم به فرم $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top Y$، محاسبه وارون $X^\top X$ با توجه به رتبه کامل نبودن ماتریس، ممکن نخواهد بود. روش‌های متفاوتی برای تحلیل این داده‌ها پیشنهاد شده است، می‌توان به رگرسیون مولفه اصلی، رگرسیون ستیغی و لاسو اشاره کرد، بطوری‌که روش منتخب با توجه به زمان، دقت و هزینه‌ای که برای محقق دارد انتخاب می‌شود. برای مطالعه بیشتر می‌توان به جولیف (۲۰۰۲)، تیبیشیرانی (۱۹۹۶)، و هورل و کنارد (۱۹۷۵) مراجعه کرد. بطوری‌که روش منتخب با توجه به زمان، دقت و هزینه‌ای که برای محقق دارد انتخاب می‌شود.

برای تحلیل داده‌های با بعد بالا، با توجه به وجود متغیرهای تبیینی فراوان، این امکان وجود دارد که بعضی از آن‌ها با متغیر پاسخ ارتباطی نداشته باشند، به همین خاطر، مؤلفه‌های اصلی به منظور کاهش ابعاد و وجود متغیرهای تبیینی که بیشترین ارتباط را با متغیر پاسخ دارند، بین روش‌های ممکن، رواج دارد. در سال‌های اخیر، یادگیری ماشین در تحلیل داده‌ها رشد چشمگیری داشته است و دانشمندان زیادی برای حل مسائل به این رویکرد متوسل شده‌اند. در میان رویکرد ها و الگوریتم‌های گوناگونی که در حوزه یادگیری ماشین وجود دارد، ماشین‌های بردار پشتیبان به عنوان یکی از مهم‌ترین و پرکاربردترین آن‌ها ابزاری قدرتمند برای طبقه‌بندی داده‌ها است، رویکرد مذکور توسط وپنیک (۱۹۹۵) پیشنهاد شده است. مدل رگرسیون بردار پشتیبان این مزیت را دارد که به دنبال خطای کمینه نیست، بلکه به دنبال خطای بهینه است. منظور از خطای بهینه، خطایی است که مدل را کاراتر کند. کنترل هواپیما بدون خلبان، آنالیز کیفیت کامپیوتر، طراحی اعضای مصنوعی، سیستم‌های مسیریابی، پیش‌بینی قیمت سهام و مواردی از این قبیل، برخی از کاربردهای این مدل است. بنابراین در این روش، نیاز به سیستم‌هایی است که توانایی یادگیری از طریق آموزش و تشخیص الگوها را دارند تا در دسته‌بندی کردن داده‌ها عملکرد مناسبی داشته باشند. برخی از محققان از الگوریتم‌های یادگیری ماشین برای افزایش عملکرد پیش‌بینی استفاده کرده‌اند (کائو و همکاران ، ۲۰۱۳؛ روزبه و همکاران ، ۱۴۰۰؛ ژایو و همکاران ، ۲۰۱۴).

تحلیل داده‌های تابعی، مبحثی در آمار است که رفتار داده‌ها در آن به‌صورت تابعی از یک متغیر دیگر است. مدل رگرسیون تابعی در زمینه‌های زیادی مانند هواشناسی، شیمی سنجی و تصویربرداری تانسور انتشار تراکتوگرافی کاربرد دارد (رامسی و سیلورمن ، ۲۰۰۵؛ فراتی و ویو ، ۲۰۰۶؛ گلدسمیت و شیپل ، ۲۰۱۴). براساس مطالعات و تحقیقات نایاک و همکاران (۲۰۱۵)، پتال و همکاران (۲۰۱۵) و آرجو و همکاران (۲۰۱۵) معیارهایی که برای ارزیابی مدل، مورد بررسی قرار می‌گیرند می‌توان به ریشه میانگین توان‌های دوم خطا[1] (RMSE) و میانگین انحراف درصد خطای مطلق[2] (MAPE)، به صورت

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(d_i - \hat{d}_i)^2}, \quad MAPE = \frac{1}{T}\sum_{i=1}^{T}\left|\frac{d_i - \hat{d}_i}{d_i}\right|$$

[1] Root of mean squared error
[2] Mean absolute percentage error

اشاره کرد، که در آن d_i مقدار واقعی داده‌ها، $\hat{d}_i$ مقدار برازش شده و T تعداد کل نمونه‌های آزمون است. بدیهی است که روش فوق با معیار انتخابی کمتر، مناسب‌تر است. در این مقاله، به برخی از روش‌ها برای مدل‌سازی روی داده‌ها بعد بالا اشاره شده است و در آخر با اشاره به دو مثال از این نوع داده‌ها، به تحلیل، انجام رگرسیون، برآورد متغیر پاسخ و انتخاب مدل مناسب با توجه به معیارهای بیان شده، پرداخته می‌شود.

۲ مدل رگرسیون تابعی

امروزه با توجه به گسترش نوع داده‌ها، نوآوری‌های فناوری اخیر در جمع‌آوری داده‌ها، ذخیره‌سازی داده‌ها، داده‌های تابعی رایج هستن (ژانگ و همکاران ، ۲۰۲۱). این مدل در بسیاری از زمینه‌های علمی مورد استفاده قرار می‌گیرد. برای مثال، دانشمندان علوم اعصاب به الگوهای اتصال تابعی بین سیگنال‌ها در مناطق مختلف مغز که در طول زمان در تصویربرداری تشدید مغناطیسی اندازه‌گیری می‌شوند، علاقه‌مند هستند (میا و همکاران ، ۲۰۲۲). برای تحلیل این گونه داده‌ها، در ابتدا، داده‌های گسسته را با توجه به روش‌های زیر به پیوسته می‌توان تبدیل کرد.

۲.۱ هموارسازی داده‌های تابعی

نخست داده‌های گسسته را باید به داده‌های پیوسته تبدیل کرد، می‌توان با برآورد یک منحنی یا یک خط راست با استفاده از روش هموارسازی، پایه‌هایی نظیر فوریه برای داده‌های دارای دوره تناوب، اسپلاین برای سایر داده‌ها و مواردی از این قبیل، این امر را محقق نمود. به طور کلی نحوه قرارگیری داده‌ها، اگر به‌صورت ترکیب خطی باشند، در این صورت:

$$Y_i = \sum_{j=1}^{k} c_j \phi_j(t_i) + \epsilon_i = f(t_i) + \epsilon_i \qquad (1)$$

که در آن تابع f ترکیب خطی از ضرایب (c_j) و تابع پایه (ϕ_j) است. همان‌طور که هر بردار در فضای برداری را می‌توان به عنوان یک ترکیب خطی از بردارهای پایه نشان داد، هر تابع پیوسته در فضای تابعی را نیز می‌توان به‌صورت ترکیبی خطی از توابع پایه نوشت. برای توابع پایه (ϕ_j) دو حالت در نظر گرفته می‌شود.

الف- تابع‌های تقریب‌زن پایه‌های فوریه اکثرا برای داده‌هایی کاربرد دارند که نوسانی بوده و دارای دوره تناوب باشند، همانند داده‌های آب و هوا که معمولا در زمستان سرد و در تابستان گرم هستند. پایه‌های فوریه به‌صورت $\{1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \ldots, \sin(m\omega t), \cos(m\omega t)\}$ معرفی می‌شوند. در تابع پایه فوق به ω دوره نوسان می‌گویند که برابر $\omega = \frac{2\pi}{p}$ است و p دوره تکرارشونده است. برای مثال، دوره تکرارشونده برای داده‌های آب و هوا برابر با ۳۶۵ است.

ب- اسپلاین‌ها همانند تابع‌های چندجمله‌ای هستند که ابتدا داده‌های گسسته را به چند قسمت مساوی تقسیم می‌کنند

و سپس به دنبال بهترین منحنی برای برازش به هر قسمت است. اگر درجه آن صفر باشد، با یک خط عمودی و افقی به برآورد پرداخته و اگر درجه‌ی آن یک باشد، به‌صورت خطی و درجه‌های بالاتر را به‌صورت منحنی برآورد می‌کند. در ضمن قسمت‌هایی از منحنی که در محل پیوستن به هم هستند را می‌توان هموار کرد، نقاطی که در قسمت اتصال قرار می‌گیرند، گره نامیده می‌شوند. برای انتخاب گره‌ها، اگر تعداد آن‌ها زیاد باشد، اریبی کم و واریانس زیاد سبب ناهمواری نمودار می‌شود. نکته قابل ذکر این است که باید در هر گره حداقل یک داده وجود داشته باشد. به‌عنوان مثال‌هایی از برخی دیگر از توابع پایه می‌توان به توابع پایه ثابت، توانی، نمایی و مواردی از این قبیل اشاره کرد.

تابع رگرسیون را می‌توان به‌صورت

$$Y_i = f(t_i) + \epsilon_i, \quad i = ۱, \ldots, n$$

در نظر گرفت که در آن خطاها مستقل و دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^۲$ هستند. با توجه به توابع پایه، برآورد $f(t_i)$ به صورت $\hat{f}(t_i) = \sum_{j=۱}^{p} c_j \phi_j(t_i)$ است، که در آن $\phi_j(t)$ تابع پایه و وابسته به نوع داده‌ها و c_j ها ضرایب هستند. برای برآورد ضرایب تابعی می‌بایست مجموع توان دوم خطاها را به‌صورت

$$\boldsymbol{H(c)} = \sum_{i=۱}^{n} (Y_i - f(t_i))^۲ = \sum_{i=۱}^{n} (Y_i - \sum_{j=۱}^{p} c_j \phi_j(t_i))^۲ \qquad (۲)$$

مینیمم کرد و همچنین می‌توان معادله فوق را به‌صورت ماتریسی $\boldsymbol{H(c)} = \boldsymbol{(Y - \Phi c)^\top (Y - \Phi c)}$ نوشت. با مشتق‌گیری و برابر صفر قرار دادن آن مقدار $\hat{\boldsymbol{c}}$ برابر $\boldsymbol{(\Phi^\top \Phi)^{-۱} \Phi^\top Y}$ خواهد شد. حال با توجه به برآورد به‌دست آمده مقدار $\hat{f}(\boldsymbol{t})$ برابر $\hat{\boldsymbol{c}}^\top \boldsymbol{\Phi}$ است، بنابراین $\hat{\boldsymbol{Y}}$ به‌صورت $\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{\Phi(\Phi^\top \Phi)^{-۱} \Phi}}_{\boldsymbol{S}} \boldsymbol{Y} = \boldsymbol{SY}$ است، که در آن $\boldsymbol{S}$ ماتریس هموارساز گفته می‌شود. در عبارت بالا مقدار $\hat{\boldsymbol{Y}}$ ضریبی از مقدار $\boldsymbol{Y}$ است، یعنی با هموارکردن داده‌ها توسط $\boldsymbol{S}$ مقدار $\hat{\boldsymbol{Y}}$ به‌دست خواهد آمد. انتخاب تعداد توابع پایه، موضوع با اهمیتی است، چرا که اگر تعداد این توابع کم باشند نشان‌دهنده اریبی زیاد و واریانس کم است و زیاد بودن تعداد توابع نشان‌دهنده اریبی کم و واریانس زیاد است که منجر به بیش برازشی مدل می‌شود.

۳ رگرسیون بردار پشتیبان

ماشین‌های بردار پشتیبان [۱] در طبقه‌بندی بسیار قوی هستند اما در رگرسیون شناخته شده نیستند. رگرسیون بردار پشتیبان [۲] یک حالت از ماشین‌های بردار پشتیبان است که به جای گرفتن مقادیر گسسته ۱− و ۱ در متغیرهای

[1] Support vector machines
[2] Support vector regression

پاسخ، مقادیر پیوسته می‌گیرد. در ماشین‌های بردار پشتیبان هر چه تعداد داده کمتری درون حاشیه قرار گیرد، خط جدا کننده مناسب‌تر است، اما در رگرسیون بردار پشتیبان با در نظر گرفتن حاشیه‌ها، هرچه تعداد داده بیشتری درون حاشیه قرار بگیرد، مدل مناسب‌تر می‌شود. هدف این بخش، ساخت مدل روی داده‌های $\{x_k, y_k\}_{k=1}^{N}$ با استفاده از رگرسیون بردار پشتیبان است که مقدار متغیر پاسخ آن پیوسته است. رگرسیون‌های ستیغی و لاسو با اضافه کردن یک پارامتر جریمه اضافی با هدف به حداقل رساندن پیچیدگی و یا کاهش تعداد ویژگی‌های مورد استفاده در مدل نهایی استفاده می‌شوند (روزبه و امینی ، ۱۳۹۸؛ روزبه و معنوی ، ۱۳۹۹)، اما در روش رگرسیون بردار پشتیبان، خطا کمینه نمی‌شود بلکه یك انعطاف پذیری وجود دارد که بتوان خطا را تغییر داد تا مدل نهایی کاراتر شود. انتخاب خطای بهینه توسط اعتبارسنجی متقابل انجام می‌شود (امینی و روزبه ، ۲۰۱۵؛ روزبه ، ۲۰۱۸).

۳.۱ عدم وجود مرز خطی در مدل رگرسیون بردار پشتیبان و حل آن

اگر مرز خطی بین داده‌ها وجود نداشته باشد، باید داده‌ها را به فضایی جدید برده و در آن فضا برای داده‌ها مرز خطی پیدا کرد. در همه‌ی مسائل فوق باید x را به $\mathbf{\Phi}(x)$ تبدیل کرد، اما چون همه داده‌ها وارد فضای جدید می‌شوند، لذا محاسبه ضرب داخلی $\mathbf{\Phi}(x)\mathbf{\Phi}(x)^\top$ بسیار طولانی است، بنابراین راه حلی معرفی می‌شود تا داده‌ها را بدون اینکه به فضای جدید تغییر داد، ضرب داخلی را بتوان حساب کرد. یکی از این راه ها استفاده از تابع هسته [1] است. چهار هسته معروف ماشین‌های یادگیری پشتیبان عبارتند از:

۱- هسته خطی[2]: ساده‌ترین تابع هسته است که حاصل ضرب داخلی $< x, y >$ به‌علاوه یک مقدار ثابت اختیاری c است.

$$k(x, y) = x^\top y + c$$

اگر داده‌ها را بتوان با یک خط از هم جدا کرد، استفاده از این روش سودمند است.

۲- هسته چندجمله‌ای[3]: هسته چندجمله‌ای زمانیکه کلیه داده‌های آموزش نرمال شده‌اند، مناسب‌تر عمل می‌کند. فرم هسته به‌صورت

$$k(x, y) = (\alpha x^\top y + c)^d$$

معرفی می‌شود. در هسته بالا پارامترهای c و α و درجه چندجمله‌ای d قابل تنظیم است.

۳- هسته گاوسی[4]: نمونه‌ایی از تابع شعاعی است که هسته آن به‌صورت:

$$k(x, y) = \exp\left(\frac{\|x - y\|^2}{2\sigma^2}\right)$$

[1] Kernel function
[2] Linear kerne
[3] Polynomial kernel
[4] Radial kernel

تعریف می‌شود. پارامتر $\sigma > ۰$ قابل تنظیم بوده و نقش عمده‌ایی در همواری هسته گاوسی دارد.

۴- هسته سیگموییید[1]: به‌عنوان هسته چند لایه پرسپترون [2] شناخته می‌شود. هسته سیگموییید از قسمت شبکه عصبی می‌آید و جایی که تابع سیگموئید دو قطبی است، اغلب به عنوان تابعی از فعال‌سازی نورون‌های مصنوعی استفاده می‌شود.

$$k(x, y) = \tanh(\alpha x^\top y + c)$$

دو پارامتر قابل تنظیم، شیب (α) و عرض از مبدا (c) در هسته سیگموییید وجود دارد.

مثال ۱. نتایج رگرسیون بردار پشتیبان برای داده‌های دو بعدی شبیه‌سازی شده با چهار هسته بالا را می‌توان در شکل ۱ مشاهده کرد. با توجه به نوع داده‌ها می‌توان نتیجه گرفت مدل با هسته شعاعی عملکرد مناسبی داشته است.

مثال ۲. برای بیان مثالی دیگر از این نوع مدل رگرسیون، می‌توان به داده‌های واقعی *faithful* از نرم افزار R مراجعه کرد. داده‌ها شامل دو بعد، زمان انتظار بین فوران‌ها و مدت فوران برای آبفشانی در پارک ملی یلوستون، وایومینگ، ایالات متحده و در سال ۱۹۹۰ میلادی است. با مدل‌سازی رگرسیون بردار پشتیبان با با هسته خطی مقدار مجذور همبستگی عدد ۰٫۸۱، هسته چندجمله‌ای عدد ۰٫۷۱، هسته سیگموییید عدد ۰٫۰۶ و هسته شعاعی عدد ۰٫۸۳ حاصل شده است. در مورد این مدل رگرسیون، هسته‌های سیگموییید عملکرد مناسبی نداشته اما سایر هسته‌ها نتایج رضایت بخش بوده است. نمودار حاصل داده‌ها و رگرسیون بردار پشتیبان را می‌توان در شکل ۲ مشاهد کرد.

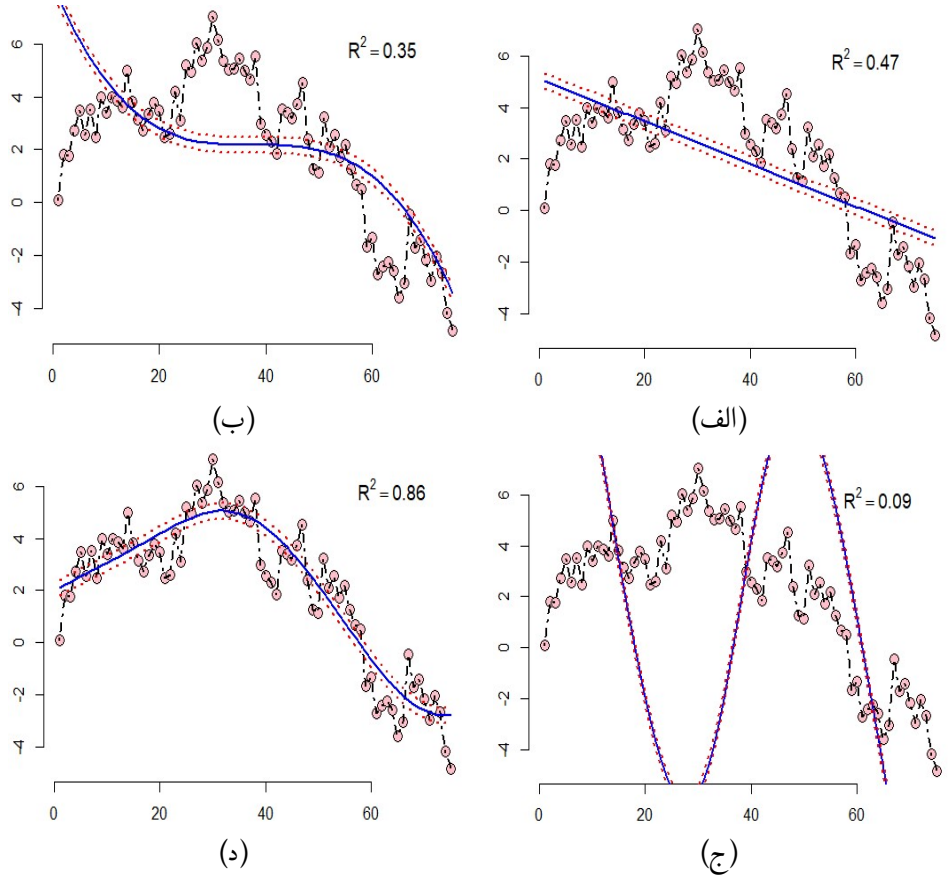
بنابر آنچه گفته شد، تکنیک رگرسیون بردار پشتیبان برای ساخت مدل به توابع هسته متکی است. از طرفی انتخاب اینکه کدام هسته با توجه به کدام داده مناسب‌تر بوده و عملکرد بهتری دارد، دشوار است و نیاز به حل مسائل بهینه‌سازی دارد.

۴ تحلیل داده‌های ریبوفلاوین

داده‌هایی که در این بخش مورد تحلیل واقع می‌شوند مربوط به تولید ریبوفلاوین (ویتامین B2) در بدن انسان هستند که مثال خوبی از داده‌های با بعد بالا محسوب می‌شوند. داده‌ها از نرم افزار R و کتاب‌خانه hdi گرفته شده است. هدف مدل‌سازی داده‌های مذکور و مقایسه مدل‌های بیان شده است. این مجموعه داده دارای ۴۰۸۸ متغیر تبیینی بوده که هر کدام نشان دهنده لگاریتم سطح ژن‌ها است. ابتدا متغیرها در داده‌های مذکور با توجه به تعداد توابع پایه بهینه شده، به منحنی تبدیل می‌شوند که منحنی‌ها را می‌توان در شکل ۳ مشاهده کرد. بنابراین، ملاحظه می‌شود:

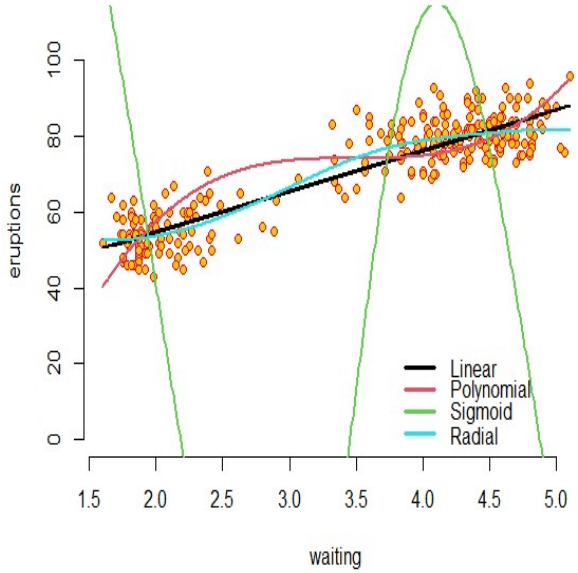
$$Y_i = \sum_{i=۱}^{۴۰۸۸} X_i(t)\beta_i(t) + \epsilon_i \qquad (۳)$$

[1] Sigmoid kernel
[2] Perceptron



شکل ۱. مدل رگرسیون بردار پشتبان با هسته الف: خطی، ب: چندجمله‌ای، ج: سیگموید، د: شعاعی

در معادله فوق Y_i لگاریتم نرخ تولید ریبوفلاوین، $X_i(t)$ بیانگر لگاریتم سطح هر ژن و $\beta_i(t)$ ضرایب تابعی مدل است. با استفاده از نمودار، بازو رگرسیون مولفه اصلی تابعی که در شکل ۴ نمایش داده شده است، تعداد مولفه‌های اصلی مورد نیاز برابر با ۱۲ مولفه است که ۸۱ درصد از واریانس کل را تبیین می‌کنند. نمودار ضرایب تابعی مدل برآورد شده در شکل ۵ مشخص و ضرایب ۱۲ مولفه منتخب به ترتیب برابر ۰٫۰۴۶۰، ۰٫۰۲۵۷−، ۰٫۰۵۱۰−، ۰٫۰۰۹۴−، ۰٫۱۵۱۶، ۰٫۰۲۶۵، ۰٫۰۳۸۰، ۵٫۵۴۰۷، ۰٫۰۴۷۹، ۰٫۰۶۱۵، ۰٫۰۰۸۳ و ۰٫۰۱۸۶ و مقدار برآورد عرض از مبدا آن برابر ۷٫۱۵۹۴ است. برای بررسی اعتبار مدل برآورد شده، طبق شکل ۶، با توجه به نمودار مانده‌ها و مانده‌های استاندارد شده، می‌توان به مناسب بودن مدل برازش شده را پی برد. در مدل رگرسیون ستیغی و لاسو، پارامتر جریمه بهینه به ترتیب برابر ۰٫۰۳۳۵۶۵۸ و ۶٫۲۸۹۶۱۸ به دست آمد. همچنین، نمودار اعتبارسنجی متقابل در برابر پارامتر جریمه در شکل ۷ رسم شد. در ادامه، با استفاده از رگرسیون بردار پشتیبان با چهار هسته (خطی، چندجمله‌ای، شعاعی و

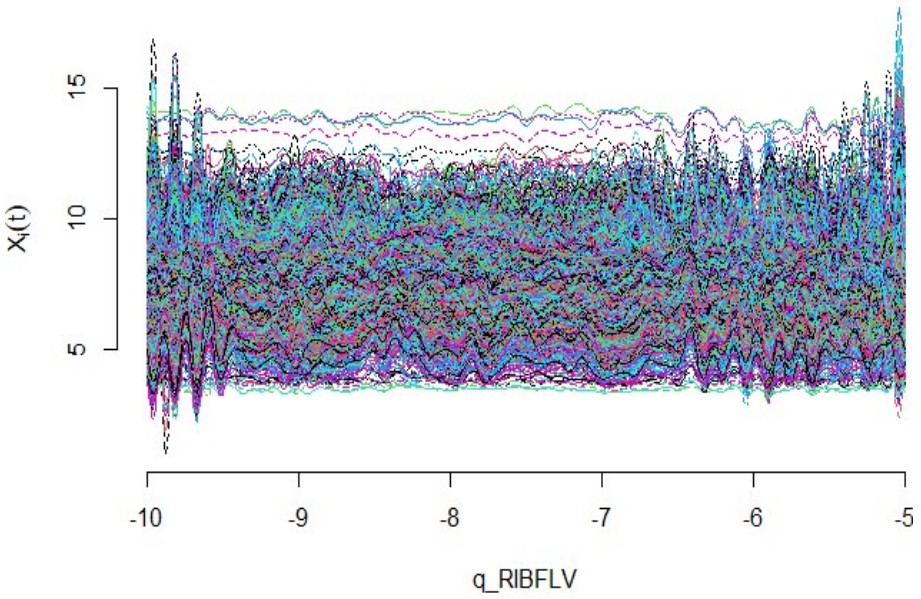


شکل ۲. مدل رگرسیون بردار پشتیبان با چهار هسته

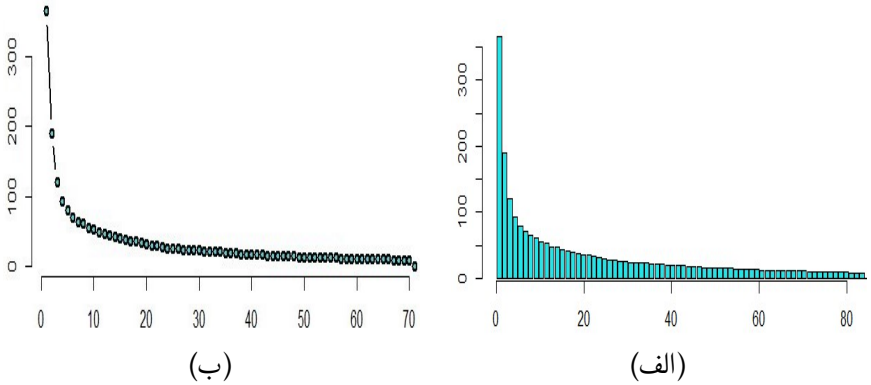
سیگموید) به مدل‌سازی و تحلیل داده‌های ریبوفلاوین پرداخته می‌شود. در این صورت:

$$Y_i = w_\circ + \sum_{i=۱}^{۴\circ۸۸} w_i X_i + \epsilon_i. \qquad (۴)$$

که در آن Y_i لگاریتم نرخ تولید ریبوفلاوین، X_i بیانگر لگاریتم سطح هر ژن و w_i ضرایب مدل رگرسیون بردار پشتیبان است. با توجه به هسته‌های بیان شده، در شکل ۸، که بیانگر رسم متغیر پاسخ در مقابل مقادیر برازش شده است، واضح است رگرسیون بردار پشتیبان با هسته خطی و سیگموید نتیجه بهتری در مقایسه با سایر هسته‌ها از خود نشان داده است. در مدل رگرسیون بردار پشتیبان با هسته خطی، بر اساس معیار اعتبارسنجی متقابل، همان‌طور که در شکل ۹ مشخص است، خطای بهینه برابر ۰/۱۴ و پارامترهای γ و c به ترتیب برابر ۱ و ۱۰ است. نتایج تمام مدل‌های پیاده شده بر روی داده‌های ریبوفلاوین در جدول ۱ گزارش شده است. نتایج به دست آمده گویای این مطلب است که رگرسیون بردار پشتیبان با هسته خطی و پارامترهای بهینه با توجه به سه معیار معرفی شده، کاراتر از سایر مدل‌ها است. همچنین می‌توان گفت که رگرسیون ستیغی هم نتایج قابل قبولی برای مدل‌سازی داده‌ها داشته است. لازم به توضیح است که رگرسیون مولفه اصلی تابعی و رگرسیون بردار پشتیبان با هسته چندجمله‌ای بر روی این داده‌ها نتایج قابل ملاحظه‌ای در مقایسه با سایر مدل‌ها نداشته است.



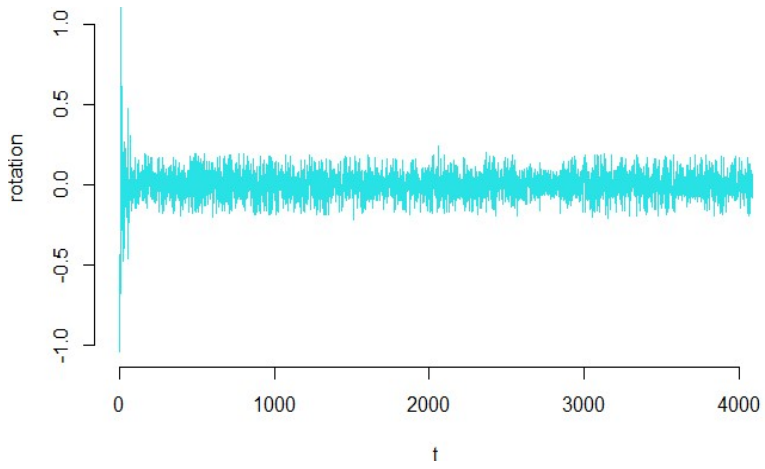
شکل ۳. منحنی‌های مربوط به داده‌های ریبوفلاوین



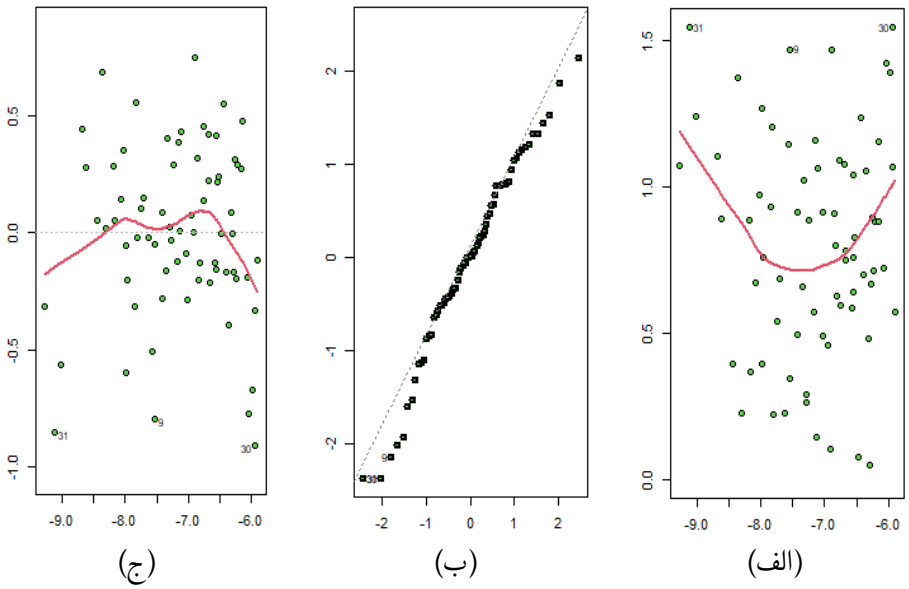
شکل ۴. الف: نمودار ستونی برای توضیح واریانس، ب: نمودار بازو برای داده‌های ریبوفلاوین

۵ مطالعه شبیه‌سازی

در این بخش مطالعات شبیه‌سازی مونت کارلو برای بررسی مدل‌هایارائه شده انجام می‌شود. با توجه به بزرگ بودن بعد داده‌ها، همان‌طور که در مقدمه نیز گفته شد، باید تعداد متغیرهای تبینی از تعداد مشاهدات بیشتر باشد

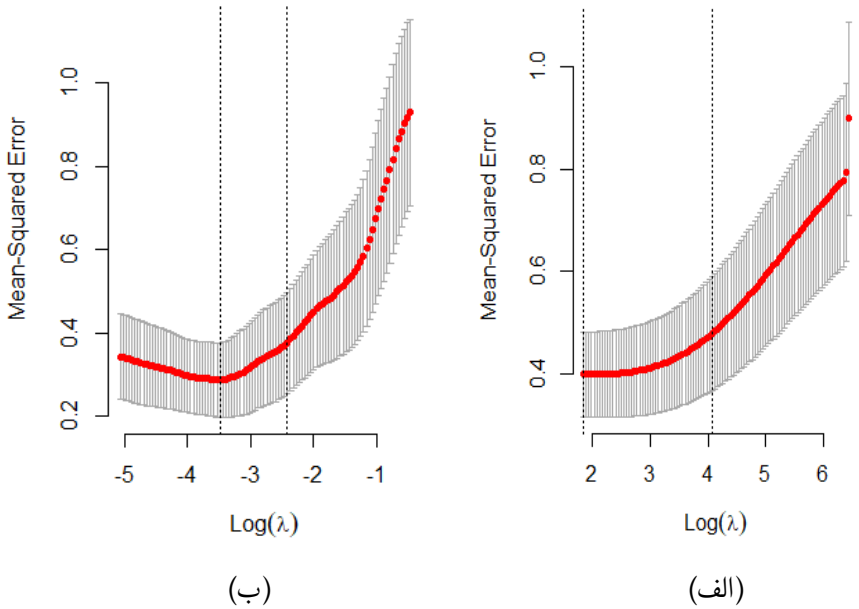


شکل ۵. نمودار ضرایب تابعی در داده‌های ریبوفلاوین



شکل ۶. انمودار لف: باقیمانده استاندارد، ب: چندک-چندک، ج: باقیمانده داده‌های ریبوفلاوین

$(p > n)$، لذا متغیرهای تبیینی با ایجاد ساختار وابستگی از مدل زیر با $n = ۲۰۰$ و $p = ۵۴۰$ شبیه سازی



شکل ۷. نمودار اعتبار سنجی برای یافتن پارامتر λ بهینه در داده‌های ریبوفلاوین، الف: جریمه لاسو، ب: جریمه ستیغی

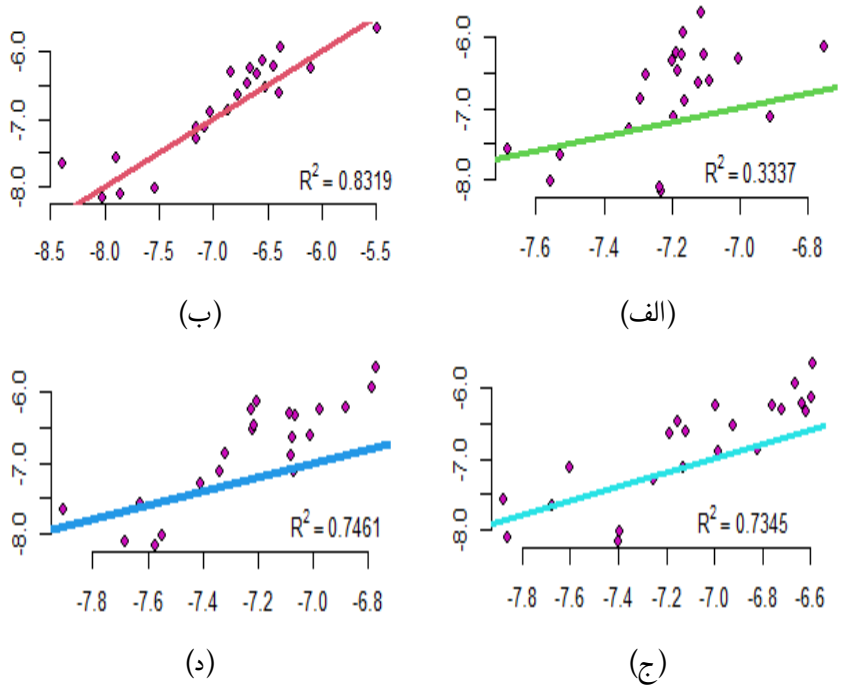
جدول ۱. جدول مقایسه بین مدل‌های رگرسیون برای داده‌های ریبوفلاوین

مدل رگرسیون	معیار		
	مجذور همبستگی	میانگین توان دوم خطا	میانگین انحراف درصد خطای مطلق
مؤلفه اصلی تابعی	۰٫۶۸۶۳	۰٫۴۲۳۸	۰٫۰۵۳۵
ستیغی	۰٫۷۸۴۸	۰٫۴۰۲۵	۰٫۰۴۹۸
لاسو	۰٫۷۶۱۷	۰٫۵۳۶۵	۰٫۰۷۳۳
بردار پشتیبان با هسته خطی	۰٫۸۳۱۹	۰٫۳۰۵۶	۰٫۰۳۵۲
بردار پشتیبان با هسته چندجمله‌ای	۰٫۳۳۳۷	۰٫۷۱۷۰	۰٫۰۸۵۴
بردار پشتیبان با هسته سیگموید	۰٫۹۴۴۲	۲۹٫۸۰۳۳	۰٫۱۵۸۳
بردار پشتیبان با هسته شعاعی	۰٫۷۴۶۱	۰٫۶۲۱۷	۰٫۰۷۵۴
بردار پشتیبان تعمیم یافته	۰٫۸۳۶۳	۰٫۳۰۷۱	۰٫۰۳۳۹

می‌شود (مک دونالد و گالارنیو ، ۱۹۷۵):

$$x_{ij} = (1-\rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip}, \qquad i = 1,\ldots,n,\ j = 1,\ldots,p \tag{۵}$$

به‌طوری که در این مدل z_{ij} اعداد تصادفی از توزیع نرمال استاندارد مستقل و ρ^2 همبستگی بین هر دو متغیر تبینی را نشان می‌دهد که در این مطالعه برابر ۰٫۹ در نظر گرفته شده است. بنابراین، متغیر پاسخ از رابطه $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ بدست می‌آید، که در آن مقادیر β_i برای $i = 1,\ldots,0.4p$ از توزیع نرمال استاندارد و برای $i > 0.4p$ برابر صفر

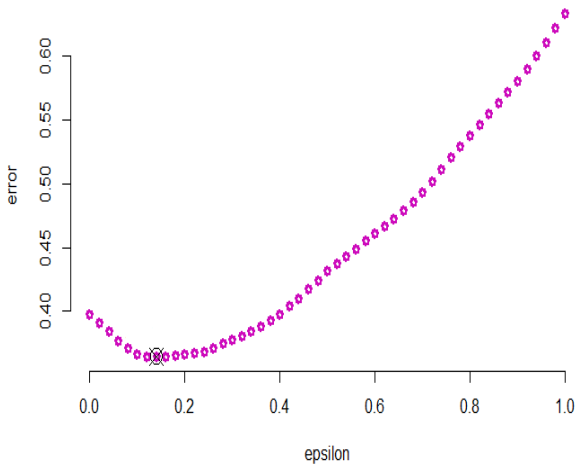


شکل ۸. نمودار مقادیر واقعی در مقابل مقادیر برازش شده با هسته الف: چندجمله‌ای، ب: خطی، ج: سیگموید، د: شعاعی در داده‌های ریبوفلاوین

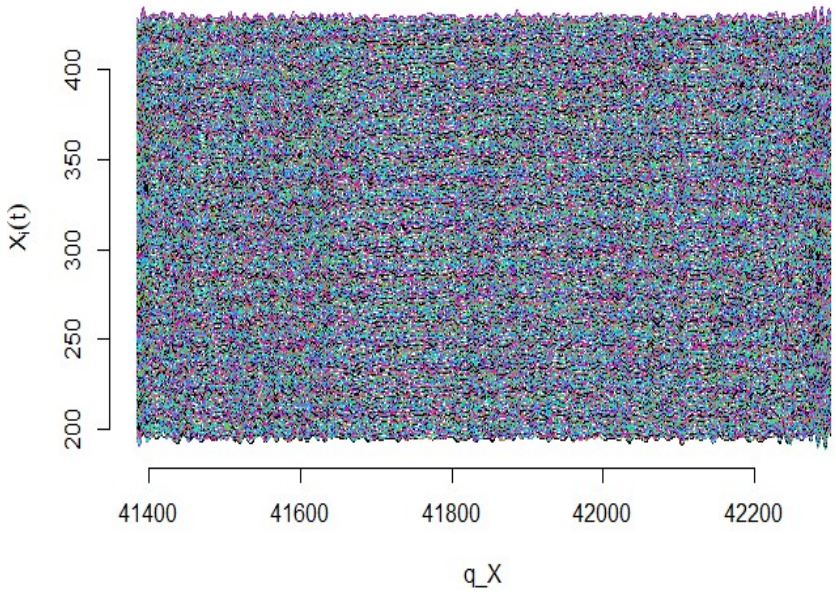
در نظر گرفته می‌شوند. همچنین مقادیر خطاها یعنی ϵ_i ها بطور تصادفی و مستقل از توزیع نرمال با میانگین صفر و واریانس ۱٫۴۴ تولید می‌شود.

اکنون با توجه به نوع داده‌ها ابتدا به برآورد منحنی‌های تابعی مربوط به متغیرهای تبینی پرداخته می‌شود، همان‌طورکه در شکل ۱۰ مشاهده می‌شود، متغیرهای تبیینی شبیه‌سازی شده با استفاده از تابع پایه بی-اسپلاین تبدیل به منحنی می‌شوند. با استفاده از رگرسیون مؤلفه اصلی، تعداد منحنی‌های مورد نیاز که بتواند اطلاعات قابل ملاحظه از کل اطلاعات موجود در داده‌ها را در بر داشته باشد انتخاب می‌کنیم. بدین منظور، نمودار بازو که در شکل ۱۱ رسم شده است، تعداد هشت مؤلفه اصلی با ۷۰ درصد از تغییرات کل داده‌ها، را پیشنهاد می‌کند. ضرایب تابعی برآورد شده در شکل ۱۲ بیانگر این موضوع است که همواری آن با توجه به نوع داده‌ها مناسب است. ضرایب مؤلفه‌های منتخب به ترتیب برابر ۲٫۱۸، ۸٫۷۵، ۱۰٫۵۳، ۹٫۸۱، ۷٫۸۸، ۶٫۱۰−، ۲٫۳۹−، و ۷٫۲۴− است.

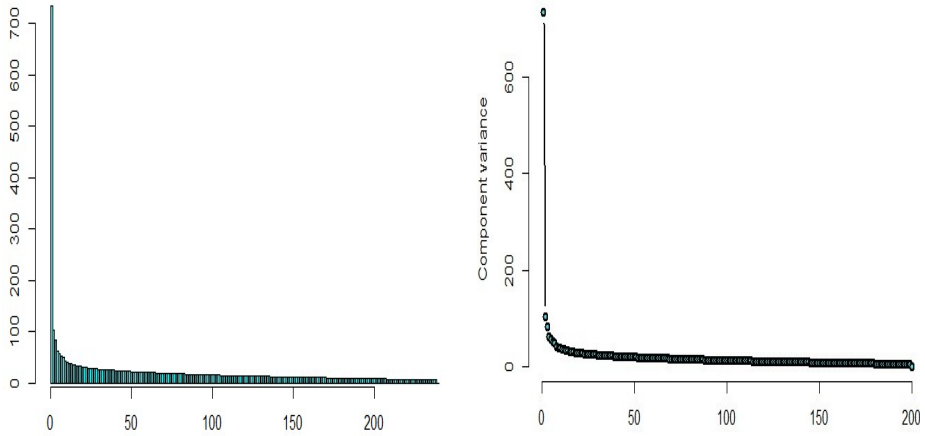
نتایج مدل رگرسیون مؤلفه اصلی تابعی، طبق شکل ۱۳، با توجه به عدم پیروی مانده‌ها از الگویی خاص و نیز قرارگیری مانده‌های استاندارد شده در محدوده ۲ و ۲−، برای داده‌های بعد بالا قابل قبول است (شیتر ، ۲۰۰۹). همچنین مدل‌سازی داده‌ها با رویکردهای رگرسیون لاسو و ستیغی با پارامتر جریمه بهینه برای هرکدام به ترتیب



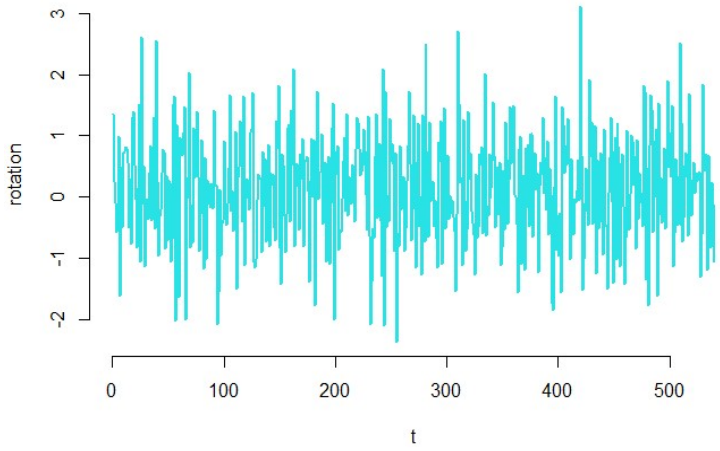
شکل ۹. نمودار اعتبارسنجی متقابل برای یافتن خطای بهینه در داده‌های ریبوفلاوین



شکل ۱۰. منحنی‌های مربوط به داده‌های شبیه‌سازی شده



شکل ۱۱. الف: نمودار ستونی برای توضیح واریانس، ب: نمودار بازو برای داده‌های شبیه‌سازی شده

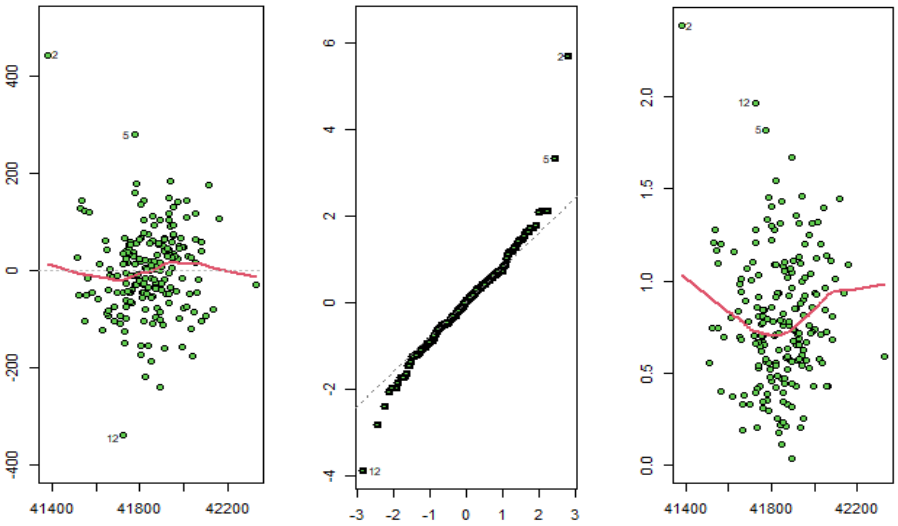


شکل ۱۲. نمودار ضرایب تابعی داده‌های شبیه‌سازی شده

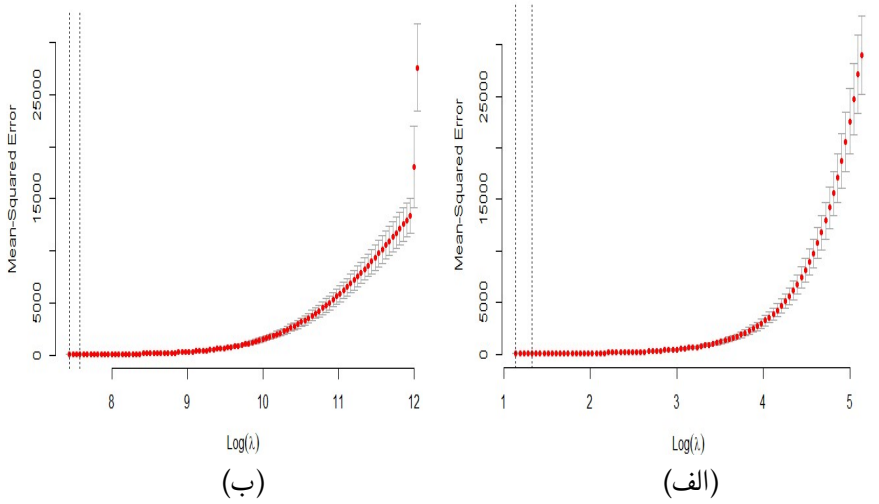
۳۱۷۰۵٫۳۰ و ۳٫۱۲ است که در شکل ۱۴ نمایش داده شده است. مجذور همبستگی بین مقادیر برآورد شده و مقادیر واقعی برای رگرسیون لاسو و ستیغی به ترتیب ۰٫۹۹۸۳ و ۰٫۹۹۸۱ به دست آمده است.

در ادامه به مدل‌سازی مجدد داده‌های شبیه‌سازی شده با استفاده از رگرسیون بردار پشتیبان

$$Y_i = w_0 + \sum_{i=1} w_i X_i + \epsilon_i, \qquad i = 1, \ldots, p \tag{۶}$$



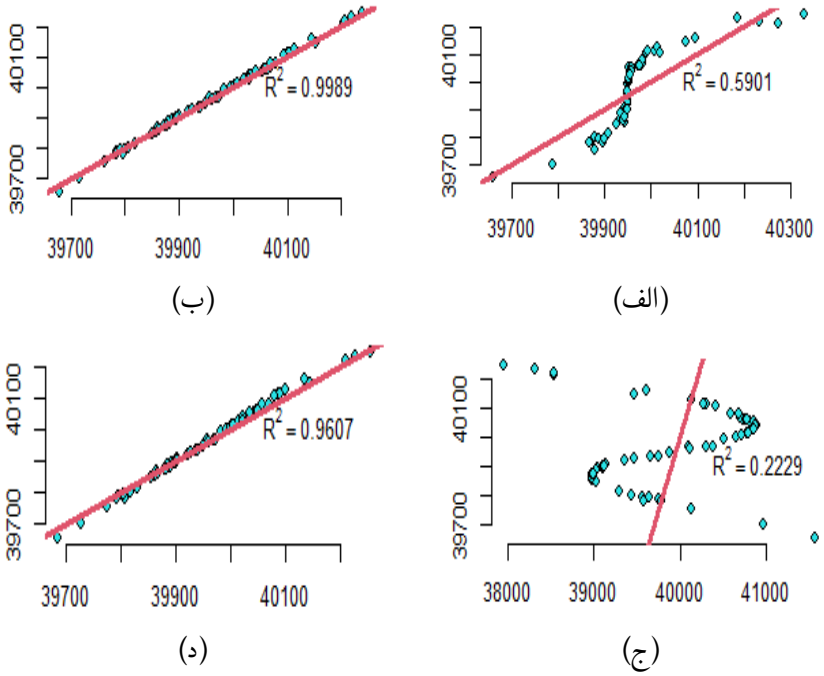
شکل ۱۳. الف: نمودار باقیمانده استاندارد، ب: نمودار چندک-چندک، ج: نمودار باقیمانده در داده‌های شبیه‌سازی شده



شکل ۱۴. نمودار اعتبار سنجی برای یافتن پارامتر λ بهینه در داده‌های شبیه‌سازی شده، الف: جریمه لاسو، ب: جریمه ستیغی

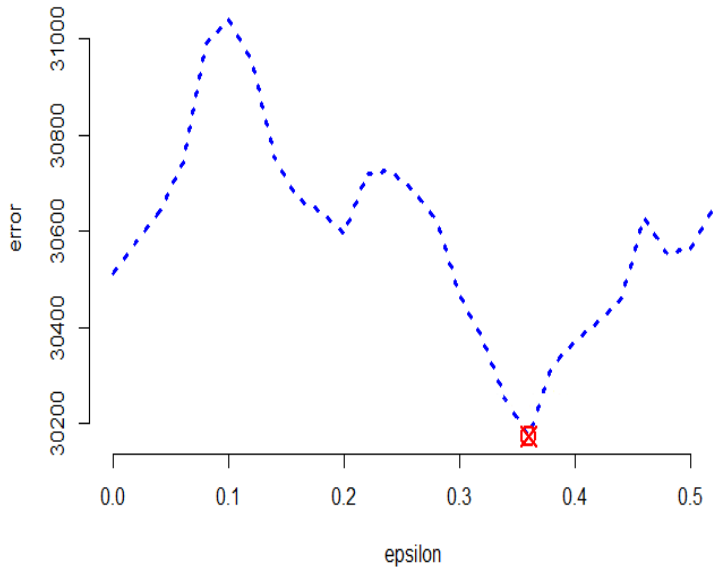
پرداخته می‌شود، که در آن w_i ضرایب مدل رگرسیون بردار پشتیبان هستند. با استفاده از چهار هسته معرفی شده، مدل‌سازی انجام شده و نتایج در شکل ۱۵ گزارش شده است. همانطورکه در این شکل دیده می‌شود، هسته‌های خطی،

چندجمله‌ای، سیگموید و شعاعی دارای مجذور همبستگی بین مقادیر واقعی و برازش شده به ترتیب برابر ۰٫۹۹۸۹، ۰٫۵۹۰۱، ۰٫۲۲۲۹ و ۰٫۹۶۰۷ هستند. لذا، در بین هسته‌های موجود با توجه به معیار مجذور هبستگی، هسته خطی و شعاعی نتایج بسیار مناسبی داشته‌اند.



شکل ۱۵. نمودار مقادیر واقعی در مقابل مقادیر برازش شده با هسته الف: چندجمله‌ای، ب: خطی، ج: سیگموید، د: شعاعی در داده‌های شبیه‌سازی شده

برای مدل رگرسیون بردار پشتیبان به دنبال خطای بهینه با توجه به چهار هسته معرفی شده، می‌باشیم. در این مدل با استفاده از هسته شعاعی و در نظر گرفتن خطای بهینه به دست آمده که در شکل ۱۶ نمایش داده شده است، مقدار اعتبار سنجی متقابل برابر ۰٫۳۶ و پارامترهای γ و c به ترتیب برابر ۰ و ۰٫۱ بدست آمده‌اند. اکنون با توجه به مدل‌سازی‌های انجام شده بر روی داده‌های شبیه‌سازی شده، به ارزیابی مدل‌های پیشنهادی بر روی داده‌های جدید می‌پردازیم. با توجه به جدول ۲ با استفاده از سه معیار مجذور همبستگی بین مقادیر واقعی و برازش شده، ریشه میانگین توان‌های دوم خطا و میانگین درصد خطای مطلق، مقایسه‌ای بین مدل‌ها صورت گرفته است. با توجه به معیار مجذور هبستگی، مدل‌های رگرسیون مؤلفه اصلی، ستیغی، لاسو و رگرسیون بردار پشتیبان با هسته‌های خطی، شعاعی و رگرسیون بردار پشتیبان تعمیم یافته نتایج مناسبی به‌همراه داشته‌اند. از نظر معیار میانگین توان دوم خطا، مدل رگرسیون بردار پشتیبان تعمیم یافته کاراتر از سایر مدل‌ها بوده و همچنین بر اساس معیار میانگین انحراف درصد



شکل ۱۶. نمودار اعتبارسنجی متقابل برای یافتن خطای بهینه در داده‌های شبیه‌سازی شده

خطای مطلق، رگرسیون مولفه اصلی تابعی، لاسو، ستیغی، رگرسیون بردار پشتیبان با هسته‌های خطی و رگرسیون بردار پشتیبان تعمیم یافته عملکردی مناسب داشته‌اند. در مجموع مدل رگرسیون بردار پشیتبان تعمیم یافته نسبت به سایر مدل‌های ارائه شده عملکرد قابل قبول و مناسبی از خود نشان داده است. ضمناً یادآور می گردد که مدل‌سازی

جدول ۲. جدول مقایسه بین مدل‌های رگرسیون برای داده‌های شبیه‌سازی شده

مدل رگرسیون	معیار		
	مجذور همبستگی	میانگین توان دوم خطا	میانگین انحراف درصد خطای مطلق
مولفه اصلی تابعی	۰٫۹۷۳۸	۸۲٫۴۹۹۵	۰٫۰۰۱۵
ستیغی	۰٫۹۹۸۱	۸٫۰۶۲۷	۰٫۰۰۰۱
لاسو	۰٫۹۹۸۳	۷٫۶۳۵۶	۰٫۰۰۰۱
بردار پشتیبان با هسته خطی	۰٫۹۹۸۹	۷٫۸۵۸۴	۰٫۰۰۰۱
بردار پشتیبان با هسته چندجمله‌ای	۰٫۵۹۰۱	۲۳۷٫۱۰۱۴	۰٫۰۰۲۴
بردار پشتیبان با هسته سیگموید	۰٫۲۲۲۹	۱۱۰۲٫۸۸۶	۰٫۰۱۸۶
بردار پشتیبان با هسته شعاعی	۰٫۹۶۰۷	۳۹٫۶۰۳۲	۰٫۰۰۰۳
بردار پشتیبان تعمیم یافته	۰٫۹۹۸۴	۶٫۴۶۹۲	۰٫۰۰۰۱

های انجام شده و نمودارهای رسم شده با استفاده از نرم افزار R و کتاب‌خانه‌های e1071، fda.usc، glmnet، و hdi بوده است.

بحث و نتیجه‌گیری

تحلیل داده‌های با بعد بالا با توجه به وارون پذیر نبودن ماتریس $X^\top X$ با شیوه‌های کلاسیک امکان پذیر نیست. در این مقاله سعی شد با بررسی مدل‌های رگرسیون ستیغی، لاسو، مولفه اصلی، تابعی و رگرسیون بردار پشتیبان با چهار هسته خطی، چندجمله‌ایِ، شعاعی و سیگموییید، به تحلیل دو داده با بعدبالا (داده‌های مربوط به تولید ریبوفلاوین و شبیه‌سازی شده) و به بررسی و تعریف رگرسیون بردار پشتیبان تعمیم یافته پرداخته شد. رگرسیون بردار پشتیبان تعمیم یافته با هسته شعاعی با پارامترهای بهینه شده $\epsilon = ۰/۳۶$، $\gamma = ۰$ و $c = ۰/۱$ در مقایسه با سایر مدل‌ها با توجه به عدد بالای مجذور همبستگی و اعداد پایین آزمون‌های میانگین توان دوم خطا و میانگین انحراف درصد خطای مطلق نسبت به سایر مدل‌ها نتایج رضایت بخشی بر روی داده‌های شبیه‌سازی شده، داشته است. در مورد داده‌های مربوط به تولید ریبوفلاوین رگرسیون بردار پشتیبان تعمیم یافته با هسته خطی و با پارامترهای بهینه شده $\epsilon = ۰/۱۴$، $\gamma = ۱$ و $c = ۱۰$ نتایج مناسبی از خود نسبت به بقیه مدل‌های ارائه شده از خود نشان داد. لازم به ذکر است که مقادیر بهینه برای رگرسیون بردار پشتیبان تعمیم یافته با استفاده اعتبارسنجی متقابل به دست آمده است.

تقدیر و تشکر

نویسندگان مقاله کمال قدردانی و تشکر را از پیشنهادات ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث ارائه بهتر و افزایش سطح کیفی مقاله شده است، دارند.

مراجع

روزبه، م. و امینی، م. (۱۳۹۸)، برآوردگر استوار مرزبندی شده تعمیم‌یافته محتمل در مدل رگرسیون نیمه‌پارامتری، مجله علوم آماری ایران، ۱۳(۲)، ۴۴۱-۴۶۰.

روزبه، م.، روحی، آ.، جهادی، ف. و زال زاده، س. (۱۴۰۰)، مدل رگرسیون بردار تکیه‌گاه و مقایسه آن با رگرسیون نیم‌پارامتری، اندیشه آماری، ۲۶(۲)، ۲۱-۳۲.

روزبه، م. و معنوی، م. (۱۳۹۹)، مدل سازی سن تقویمی به روش رگرسیون ستیغی کمترین توان های دوم پیراسته، مجله علوم آماری ایران، ۱۴(۲)، ۴۰۹-۴۲۸.

Amini, M. and Roozbeh, M. (2015), Optimal Partial Ridge Estimation in Restricted Semiparametric Regression Models, *Journal of Multivariate Analysis*,**136**, 26-40.

Araújo R. D. A., Oliveira A. L. and Meira S. (2015), A Hybrid Model for High-Frequency Stock Market Forecasting, *Expert Systems with Applications*, **42** , 4081-4096.

Efron, B., and Hastie,T. (2017), Computer Age Statistical Inference, *Cambridge University Press*, Cambridge.

Ferraty, F., Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science and Business Media, New York,

Jolliffe I.T. (2002), *Principal Component Analysis*, Springer Series in Statistics, Aberdeen.

Hoerl A.E. and Kennard R.W. (1975), Ridge Regression some Simulation, *Communication in Statistics*, **4** , 4105–4123.

Goldsmith, J., Scheipl, F. (2014), Estimator Selection and Combination in Scalar-on-Function Regression, *Computational Statistics & Data Analysis*, **70**, 362-372.

Kao L. J., Chiu C. C., Lu C. J. and Yang J. L. (2013), Integration of Nonlinear Independent Component Analysis and Support Vector Regression for Stock Price Forecasting, *Neurocomputing*, **99** , 534-542.

McDonald G. C., and Galarneau D. I. (1975), A Monte Carlo Evaluation of some Ridge-Type Estimators. *Journal of the American Statistical Association*, **70(350)**, 407-416.

Miao, R., Zhang, X., and Wong, R. K. (2022), A Wavelet-Based Independence Test for Functional Data with an Application to MEG Functional Connectivity. *Journal of the American Statistical Association*, 1-14.

Nayak R. K., Mishra D. and Rath A. K. (2015), A Naïve Svm-Knn Based Stock Market Trend Rreversal Analysis for Indian Benchmark Indices, *Applied Soft Computing*, **35** , 670-680.

Patel J., Shah S., Thakkar P. and Kotecha K. (2015), Predicting Stock Market Index Using Fusion of Machine Learning Techniques,*Expert Systems with Applications*, **42** , 2162-2172.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer-Verlag, New York.

Roozbeh, M. (2018), Optimal QR-Based Estimation in Partially Linear Regression Models with Correlated Errors Using GCV Criterion, *Computational Statistics & Data Analysis*, **117**, 45-61.

Sheather, S. (2009), *A Modern Approach to Regression with R*, Springer Science and Business Media.

Tibshirani R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society.* Series B (Methodological), **58**(1) , 267-288.

Vapnik V. N. (1995), *The Nature of Statistical Learning Theory*, New York.

Xiao Y., Xiao J., Lu F. and Wang S. (2014), Ensemble Anns-Pso-Ga Approach for Day-Ahead Stock E-Exchange Prices Forecasting, *International Journal of Computational Intelligence Systems*, **7** , 272-290.

Zhang, X., Xue, W., and Wang, Q. (2021), Covariate Balancing Functional Propensity Score for Functional Treatments in Cross-Sectional Observational Studies. *Computational Statistics & Data Analysis*, **163**, 107303.