Statistical Inference

Lecturer: Abdol-Hossein Vahabie

Spring Semester 1401-1402

Marzieh Alidadi_810101236 Writing Assignment IV Deadline 1402/03/06

۱ پاسخ کوتاه

۱-۱ زیربخش ۱

رگرسیون خطی ساده تنها دارای یک متغیر x و یک متغیر y است. رگرسیون خطی چندگانه دارای یک متغیر y و دو یا بیشتر از دو متغیر x است.

۱-۲ زیربخش ۲

slope به عنوان تغییرات ایجاد شده در y به ازای هر افزایش یک واحدی در x تفسیر می‌شود.

۱-۳ زیربخش ۳

Residual analysis یک کلاس مفید از تکنیک‌ها است که برای ارزیابی goodness مدل‌ها استفاده می‌شود. این روش به ویژه در رگرسیون چندگانه مفید است، زیرا گاهی اوقات Scatter Plot برای ارزیابی تصویری در دسترس نیست.

۱-۴ زیربخش ۴

link function ها در GLM ها، رابطه‌ی پیش‌بینی‌کننده‌ی خطی را با متغیر هدف برقرار می‌کنند. انتخاب link function به دلیل این‌که رابطه‌ی بین متغیرها را تعیین می‌کند، بر تفسیر ضرایب در مدل تاثیر می‌گذارد. link function های مختلف ممکن است منجر به تفاسیر مختلفی درباره‌ی اندازه و جهت تأثیر پیش‌بینی‌کننده‌ها بر متغیر هدف شوند.

۱-۵ زیربخش ۵

رگرسیون Ridge یک تکنیک نرمال سازی است، که از آن در رگرسیون خطی برای جلوگیری از بیش‌برازش استفاده شده‌است. در این تکنیک، یک عبارت جریمه (penalty) به تابع هدف رگرسیون Ordinary Least Squares اضافه می‌شود که ضرایب را به سمت صفر کوچک می‌کند. رگرسیون Ridge با Ordinary Least Squares در این مورد تفاوت دارد که با وارد کردن انحراف (bias) به تخمین‌ها، واریانس را کاهش می‌دهد. که این مسئله، زمانی که بین متغیرهای پیش‌بینی کننده همبستگی بالایی وجود دارد یا با داده‌های با بعد بالا سروکار داریم، می‌تواند موجب بهبود عملکرد Least Squares Ordinary شود.

۱-۶ زیربخش ۶

زمانی مدل پواسون بهتر از مدل رگرسیون خطی است، که متغیر هدف، از نوع شمارش (count) یا نرخ (rate) باشد و دارای توزیع پواسون با واریانسی متناسب با میانگینش باشد. این نوع مدل در شرایطی مفید است که داده‌ها شامل رویدادهای مبتنی بر شمارش باشند.

۱-۷ زیربخش ۷

هنگامی که متغیرهای مستقل در مدل رگرسیون با یکدیگر به شدت همبستگی دارند (در صورتی که باید این متغیرها از هم مستقل باشند)، به این وضعیت Multicollinear-ity گفته می‌شود. این وضعیت باعث سخت شدن تفسیر مدل و همچنین ایجاد مشکل بیش‌برازش می‌شود.

برخی از روش‌های حل این مشکل در یک مدل رگرسیون خطی، عبارتند از:

۱. حذف برخی از متغیرهای مستقل با همبستگی بالا.

۲. ترکیب خطی متغیرهای مستقل، مانند جمع آن‌ها.

۳. رگرسیون partial least squares از تجزیه مؤلفه‌های اصلی استفاده می‌کند، تا یک مجموعه از مؤلفه‌های بدون همبستگی را برای درج در مدل ایجاد کند.

۴. رگرسیون‌های LASSO و Ridge نسخه‌های پیشرفته‌ای از تحلیل رگرسیون هستند که می توانند این مشکل را حل کنند.

۱-۸ زیربخش ۸

Heteroscedasticity یک وضعیت در رگرسیون خطی است، که در آن تنوع خطا در سراسر دامنه‌ی مقادیر متغیر مستقل نامساوی است. این موضوع می‌تواند بر روی دقت ضرایب و خطاهای استاندارد تخمین‌زده شده تأثیر بگذارد و باعث انحراف (bias) نتایج شود. یک روش معمول برای تست Heteroscedasticity رسم نمودار residual ها در مقابل مقادیر پیش‌بینی شده و جستجوی الگوها است. روش دیگر، تست Breusch-Pagan است که وجود ارتباط میان واریانس residual ها و متغیرهای مستقل را بررسی می‌کند.

۱-۹ زیربخش ۹

اصلی ترین تفاوت بین رگرسیون logistic و رگرسیون خطی این است که رگرسیون logistic برای مسائل دسته‌بندی استفاده می‌شود، در حالی که رگرسیون خطی برای پیش‌بینی مقادیر پیوسته استفاده می‌شود.

۱-۱۰ زیربخش ۱۰

Overdispersion در مدل‌های خطی عمومی زمانی رخ می‌دهد که واریانس داده‌ها از آنچه که مدل فرض می‌کند بیشتر است. برای مقابله با این موضوع، می‌توان از رویکرد quasi-likelihood یا مدل binomial negative به جای مدل پواسون استفاده کرد. راه دیگری برای مقابله با Overdispersion شامل اضافه کردن متغیرهای توضیحی اضافه است، که ممکن است تنوع اضافه را در داده‌ها توجیه کنند.

۱-۱۱ زیربخش ۱۱

نسبت‌های شانس، در مدل‌های رگرسیون logistic نشان‌دهنده‌ی تغییر در شانس وقوع یک رویداد، در ارتباط با تغییر یک واحدی در متغیر پیش‌بینی‌کننده است، در حالی که همه متغیرهای کمکی دیگر ثابت نگه داشته می‌شوند. نسبت شانس بزرگ‌تر از ۱ نشان دهنده‌ی ارتباط مثبت و نسبت شانس کم‌تر از ۱ نشان دهنده‌ی ارتباط منفی است.

۱-۱۲ زیربخش ۱۲

بله، متغیرهای دسته‌ای می‌توانند با تبدیل به متغیرهای dummy در یک مدل رگرسیون خطی یا خطی تعمیم‌یافته استفاده شوند.

۱-۱۳ زیربخش ۱۳

نرمال‌سازی، یک تکنیک است که با افزودن یک عبارت جریمه (penalty) به تابع هزینه، در یک مدل رگرسیون خطی یا خطی تعمیم‌یافته برای جلوگیری از بیش‌برازش استفاده می‌شود. این تکنیک به ساده‌تر شدن مدل و بهبود قابلیت تعمیم‌پذیری آن در مقابل داده‌های جدید کمک می‌کند.

۱-۱۴ زیربخش ۱۴

بله، outlier می‌تواند نتایج یک مدل رگرسیون خطی یا خطی تعمیم‌یافته را تحت تاثیر قرار داده و باعث پیش‌بینی‌های نادرست شود. برای رفع این مشکل، ابتدا باید outlier ها را شناسایی کرده و سپس آن‌ها را از مجموعه داده حذف کنیم یا تکنیک‌هایی مانند رگرسیون robust یا winsorization را روی آن‌ها اعمال کنیم.
برای شناسایی outlier در داده‌های خود، می‌توان با استفاده از نمودارهای پراکندگی یا box plot داده‌ها را تجسم کرده و هر مشاهده‌ای که به شدت متفاوت با سایر مشاهدات است را شناسایی کرد. همچنین، از آزمون‌های آماری مانند Z-score یا IQR برای شناسایی outlier می‌توان استفاده کرد.

۱-۱۵ زیربخش ۱۵

یک روش برای ارزیابی مطلوبیت تناسب یک مدل رگرسیون logistic با استفاده از آزمون Hosmer-Lemeshow است که احتمالات پیش‌بینی شده را با نتایج مشاهده شده مقایسه می‌کند.

۱-۱۶ زیربخش ۱۶

بله، در هر دو مدل رگرسیون خطی چندگانه و رگرسیون خطی تعمیم‌یافته، می‌توان عبارات تعاملی (interaction terms) را در نظر گرفت. استفاده از عبارات تعاملی، کمک می‌کند تا تأثیر یک متغیر پیش‌بین بر روی نتیجه را با در نظر گرفتن تأثیر متغیر دیگری که در مدل قرار دارد، مدل کنیم. تفسیر عبارات تعاملی، بر اساس بررسی نحوه‌ی تغییر رابطه بین یک متغیر پیش‌بین با نتیجه در سطوح مختلف متغیر دیگر صورت می‌گیرد.

۱-۱۷ زیربخش ۱۷

روش‌های پارامتری فرض می‌کنند که رابطه بین متغیرها با یک شکل عملکردی خاصی توصیف می‌شود و مبتنی بر این فرض، پارامترهای مدل تخمین زده می‌شوند. از سوی دیگر، روش‌های ناپارامتری هیچ فرضی درباره شکل رابطه‌ی بین متغیرها نمی‌کنند و با استفاده از تکنیک‌های انعطاف‌پذیر آن را تخمین می‌زنند.

۱-۱۸ زیربخش ۱۸

روش های متداول برای پر کردن مقادیر گم شده در داده‌های سری زمانی، شامل تفسیر خطی، پر کردن به صورت جلو و عقب، و میانگین‌گیری هستند. با این حال، برای داده‌های بازار سهام، توصیه می‌شود از روش‌های پیشرفته‌تری مانند مدل‌های پیش‌بینی سری زمانی ARIMA و سایر مدل‌های پیچیده‌تر استفاده شود که الگوها و نوسانات پیچیده داده‌های مالی را بیشتر می‌توانند درک کنند.

۱-۱۹ زیربخش ۱۹

خیر، روابط غیرخطی نمی‌توانند با استفاده از رگرسیون خطی چندگتنه یا خطی تعمیم@یافته به درستی مدل شوند. برای این کار باید از مدل‌های رگرسیون غیرخطی استفاده شود.

۱-۲۰ زیربخش ۲۰

نمودارهای ACF و PACF در تحلیل سری‌های زمانی برای شناسایی وجود و قدرت auto-correlation در داده‌ها استفاده می‌شوند. ACF اندازه‌ی رابطه‌ی بین یک سری زمانی و مقادیر دارای lag آن را اندازه‌گیری می‌کند، در حالی که PACF رابطه‌ی بین یک سری زمانی و مقادیر دارای lag آن را بعد از حذف تأثیر سایر lag ها اندازه‌گیری می‌کند. این نمودارها به ما کمک می‌کنند تا پارامترهای بهینه برای مدل‌های autoregressive و moving average را شناسایی کنیم.

۱-۲۱ زیربخش ۲۱

دو آزمون آماری که به طور معمول برای تعیین white noise بودن/نبودن یک سری زمانی استفاده می‌شوند، آزمون Ljung-Box و آزمون Augmented Dickey-Fuller (ADF) هستند.
آزمون Ljung-Box با بررسی اینکه آیا گروهی از auto-correlation ها به شدت از صفر متمایزند، auto-correlation در یک سری زمانی را بررسی می‌کند. اگر مقدار p-value کمتر از سطح معناداریِ انتخاب شده باشد، پس شواهدی برای عدم تصادفی بودن داده‌ها وجود دارد و سری زمانی یک white noise نیست.
آزمون ADF برای شناسایی وجود ریشه واحد در یک سری زمانی که نشان‌دهنده‌ی عدم استحکام و عدم تصادفی بودن است، بررسی می‌شود. اگر مقدار p-value کمتر از سطح معناداریِ انتخاب شده باشد، آنگاه فرض صفر یک ریشه واحد رد می‌شود و سری زمانی یک white noise نیست.

۱-۲۲ زیربخش ۲۲

دقت و عملکرد یک مدل رگرسیون می‌تواند با استفاده از معیارهایی مانند
(MSE) mean squared error
(RMSE) root mean squared error
(R-squared) coefficient of determination
و (MAE) mean absolute error
اندازه‌گیری شود.

۱-۲۳ زیربخش ۲۳

اصطلاح "Stationarity" در داده‌های سری زمانی به معنای این است که خصوصیات آماری داده‌ها با گذشت زمان تغییر نمی‌کنند. اگر داده‌های ما این ویژگی را نداشته باشند، باید قبل از تحلیل آن‌ها، با استفاده از تفاضل گیری و یا روش های دیگر، آن‌ها را به یک فرآیند Stationarity تبدیل کنیم.

۱-۲۴ زیربخش ۲۴

می‌توان از ANOVA در زمینه رگرسیون خطی برای آزمون قابل توجه بودن کلی مدل استفاده کرد. آماره‌ی F به دست آمده از ANOVA نسبت واریانس توضیح داده شده توسط مدل به واریانس توضیح داده نشده توسط مدل را نمایش می‌دهد. در این زمینه، متغیرها، پیش‌بینی‌کننده‌های مدل را نمایش می‌دهند و ضرایب، نمایانگر اثر هر پیش‌بینی‌کننده بر متغیر خروجی هستند.

۱-۲۵ زیربخش ۲۵

شرایط استنتاج در رگرسیون خطی عبارتند از:

۱. رابطه خطی بین متغیر وابسته (هدف) و مستقل.

۲. Homoscedasticity به معنی واریانس یکسان خطاها در تمام مقادیر متغیر مستقل.

۳. عدم وابستگی خطاها به یکدیگر.

۴. نرمال بودن توزیع خطاها.

۱-۲۶ زیربخش ۲۶

در نمودار residual اول، نقاط به طور تصادفی در اطراف خط residual=۰ پراکنده شده‌اند. می‌توان نتیجه گرفت که یک مدل خطی برای مدل سازی این داده‌ها مناسب است.

در نمودار residual دوم، نقاط به صورت U شکل در بالا و پایین خط residual=۰ قرار گرفته‌اند. به همین دلیل، نمی‌توان از یک مدل خطی برای مدل‌سازی این داده‌ها

استفاده کرد. ابتدا باید یک تبدیل مناسب روی این متغیرها انجام شود تا دارای رابطه‌ی خطی شوند، و سپس از یک مدل خطی برای مدل‌سازی آن‌ها استفاده شود.
در نمودار residual سوم، نقاط از هیچ یک از دو الگوی فوق تبعیت نمی‌کنند. و به طور کاملاً پراکنده در صفحه قرار گرفته اند و عملاً نمی‌توان آن‌ها را مدل کرد.

۱-۲۷ زیربخش ۲۷

از توزیع t با درجه آزادی n-۲ برای برآورد پارامترهای رگرسیون beta۰ و beta۱ استفاده می‌شود؛ زیرا این توزیع، عدم قطعیت در برآورد انحراف معیار نمونه را در نظر می‌گیرد و بهترین برآورد برای انحراف معیار جمعیت را فراهم می‌کند.

۱-۲۸ زیربخش ۲۸

این بازه اطمینان به معنای آن است که با ۹۵ درصد اطمینان، مقدار واقعی پارامتر beta۱ رگرسیون بین ۰.۳۴- و ۴ قرار دارد. به عبارت دیگر، اگر ما مطالعه را چندین بار تکرار کنیم و در هر بار بازه اطمینان را محاسبه کنیم، انتظار داریم که مقدار واقعی پارامتر در ۹۵ درصد این بازه‌های محاسبه شده قرار داشته باشد. این بدین معنا نیست که مقدار واقعی، با احتمال ۹۵ درصد در این بازه قرار می‌گیرد، بلکه مقدار واقعی یا در بازه قرار دارد یا نه، و سطح اطمینان ما مربوط به بازه است، نه مقدار واقعی پارامتر.

۱-۲۹ زیربخش ۲۹

AIC و BIC معیارهایی هستند که در مدل‌سازی آماری برای ارزیابی مطلوبیت تطبیق مدل با داده‌های مختلف استفاده می‌شوند. این دو معیار یک اندازه‌گیری از تعادل بین پیچیدگی مدل (تعداد پارامترها) و مطلوبیت تطبیق مدل با داده‌ها ارائه می‌دهند. AIC به مدل‌های ساده‌تر امتیاز بیشتری می‌دهد، در حالی که BIC مدل‌های پیچیده را بیشتر تنبیه می‌کند. این معیارها می‌توانند برای مقایسه مدل‌های مختلف و انتخاب مدلی که بهترین تعادل بین پیچیدگی مدل و مطلوبیت تطبیق با داده‌ها را برای مجموعه داده خاصی دارد، استفاده شوند.

۳۰-۱ زیربخش ۳۰

از روند نشان داده شده در جدول، می‌توان به این نتیجه رسید که افزایش درصد افراد چاق در ایالات متحده ممکن است ادامه یابد و تا سال ۲۰۵۰ به حدود ۶۰-۶۵ درصد برسد. اما این روش دارای محدودیت‌ها و ابهاماتی‌ست. عوامل خارجی مانند تغییرات رژیم غذایی، سبک زندگی و پیشگیری از بیماری‌ها می‌توانند این روند را تحت تأثیر قرار دهند و باعث تغییرات غیرمنتظره شوند. بنابراین، در حالی که این روش به یک تخمین خشک و ساده می‌انجامد، باید با احتیاط و به همراه محدودیت‌ها و تخمین‌های نامطمئن همراه باشد.

۲ درست یا غلط؟

۲-۱ زیربخش ۱

درست

۲-۲ زیربخش ۲

درست

۲-۳ زیربخش ۳

درست

۲-۴ زیربخش ۴

غلط

۲-۵ زیربخش ۵

درست

۲-۶ زیربخش ۶

درست

۲-۷ زیربخش ۷

درست

۲-۸ زیربخش ۸

درست

۲-۹ زیربخش ۹

درست

۲-۱۰ زیربخش ۱۰

درست

۲-۱۱ زیربخش ۱۱

درست

۲-۱۲ زیربخش ۱۲

غلط

۲-۱۳ زیربخش ۱۳

درست

۲-۱۴ زیربخش ۱۴

غلط

۲-۱۵ زیربخش ۱۵

درست

۲-۱۶ زیربخش ۱۶

درست

۲-۱۷ زیربخش ۱۷

درست

۲-۱۸ زیربخش ۱۸

درست

۲-۱۹ زیربخش ۱۹

درست

۲-۲۰ زیربخش ۲۰

درست

۲-۲۱ زیربخش ۲۱

درست

۲-۲۲ زیربخش ۲۲

درست

۲-۲۳ زیربخش ۲۳

غلط

۲-۲۴ زیربخش ۲۴

غلط

۲-۲۵ زیربخش ۲۵

درست

۳ رگرسیون خطی

۳-۱ زیربخش a

$$n = 5 \tag{۱}$$

$$\sum_{i=1}^{5} X = 35 + 40 + 45 + 50 + 55 = 225 \tag{۲}$$

$$\sum_{i=1}^{5} Y = 450 + 500 + 550 + 600 + 650 = 2750 \tag{۳}$$

$$\sum_{i=1}^{5} XY = 35(450) + 40(500) + 45(550) + 50(600) + 55(650) = 126250 \tag{۴}$$

$$\sum_{i=1}^{5} X^2 = 35^2 + 40^2 + 45^2 + 50^2 + 55^2 = 10375 \tag{۵}$$

$$\begin{aligned} a = \frac{\sum_{i=1}^{5} Y * \sum_{i=1}^{5} X^2 - \sum_{i=1}^{5} X * \sum_{i=1}^{5} XY}{n * \sum_{i=1}^{5} X^2 - (\sum_{i=1}^{5} X)^2} &= \frac{2750 * 10375 - 225 * 126250}{5 * 10375 - 225^2} \\ &= \frac{28531250 - 28406250}{51875 - 50625} = \frac{125000}{1250} = 100 \end{aligned} \tag{۶}$$

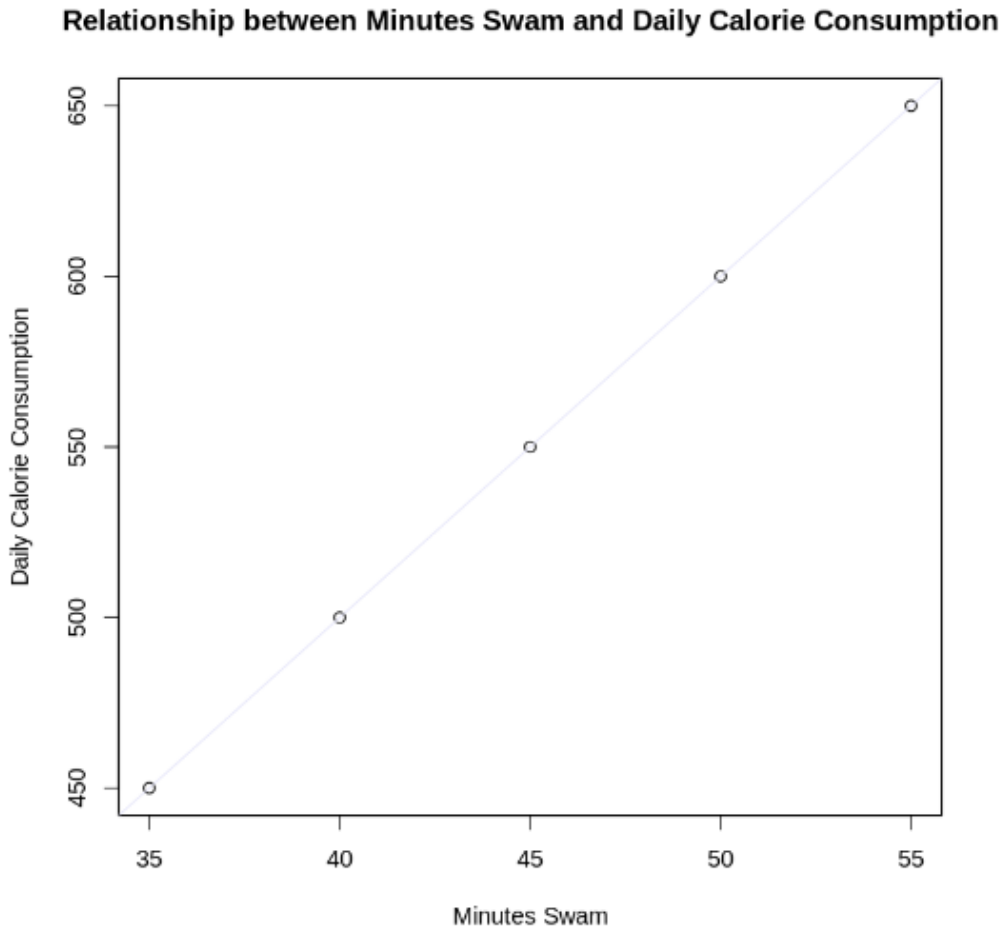
$$b = \frac{n * \sum_{i=1}^{5} XY - \sum_{i=1}^{5} X * \sum_{i=1}^{5} Y}{n * \sum_{i=1}^{5} X^2 - (\sum_{i=1}^{5} X)^2} = \frac{5 * 126250 - 225 * 2750}{5 * 10375 - 225^2} = \frac{631250 - 618750}{51875 - 50625} = \frac{12500}{1250} = 10 \quad (۷)$$

رگرسیون خطی به فرم زیر خواهد بود:

$$Y = a + bX \quad (۸)$$

$$Y = 100 + 10X$$

نمودار scatter plot نشان‌دهنده‌ی رابطه‌ی خطی بین این دو متغیر به شرح زیر است:



۲-۳ زیربخش b

میزان مصرف کالری روزانه برای افرادی که ۴۸ دقیقه در روز شنا می‌کنند، با استفاده از رگرسیون خطی تولید شده، به شرح زیر پیشبینی خواهد شد:

$$Y = 100 + 10X = 100 + 10(48) = 100 + 480 = 580 \quad (۹)$$

۳-۳ زیربخش c

مقادیر پیشبینی شده برای میزان مصرف کالری روزانه افراد با استفاده از رگرسیون خطی تولید شده، به شرح زیر است:

$$Y_1 = 100 + 10X = 100 + 10(35) = 100 + 350 = 450 \quad (۱۰)$$

$$Y_2 = 100 + 10X = 100 + 10(40) = 100 + 400 = 500 \quad (۱۱)$$

$$Y_3 = 100 + 10X = 100 + 10(45) = 100 + 450 = 550 \quad (۱۲)$$

$$Y_4 = 100 + 10X = 100 + 10(50) = 100 + 500 = 600 \quad (۱۳)$$

$$Y_5 = 100 + 10X = 100 + 10(55) = 100 + 550 = 650 \quad (۱۴)$$

مقادیر واقعی میزان مصرف کالری روزانه افراد در جدول آورده شده در سوال نشان داده شده‌است.
با استفاده از مقادیر واقعی و مقادیر پیشبینی شده برای میزان مصرف کالری روزانه‌ی افراد، مقدار RSS به فرم زیر محاسبه می‌شود:

$$RSS = ResidualSumOfSquares = \sum_{i=1}^{5}(actual_i - predicted_i)^2 = 0 \qquad (۱۵)$$

میانگین میزان مصرف کالری روزانه افراد به شرح زیر است:

$$\bar{Y} = \frac{\sum_{i=1}^{5}(actual_i)}{5} = \frac{450 + 500 + 550 + 600 + 650}{5} = \frac{2750}{5} = 550 \qquad (۱۶)$$

با استفاده از مقادیر واقعی و میانگین میزان مصرف کالری روزانه‌ی افراد، مقدار TSS به فرم زیر محاسبه می‌شود:

$$\begin{aligned} TSS = TotalSumofSquares = \sum_{i=1}^{5}(actual_i - \bar{Y})^2 = \\ (450-550)^2 + (500-550)^2 + (550-550)^2 + (600-550)^2 + (650-550)^2 \\ = (-100)^2 + (-50)^2 + (0)^2 + (50)^2 + (100)^2 \\ = 10000 + 2500 + 0 + 2500 + 10000 = 25000 \end{aligned} \qquad (۱۷)$$

با توجه به TSS و RSS محاسبه شده، مقدار coefficient of determination به شرح زیر بدست می‌آید:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{0}{25000} = 1 - 0 = 1 \qquad (۱۸)$$

مقدار بدست آمده برای R-Squared بدین شرح تفسیر می‌شود که مدل رگرسیون

خطی تولید شده، به خوبی، خروجی (Y) را پیشبینی می‌کند.

۴ رگرسیون خطی چندگانه

۴-۱ زیربخش a

فرض صفر و فرض جایگزین:
-فرض صفر، این را بیان می‌کند که ضرایب استفاده شده در مدل رگرسیون خطی، همگی برابر ۰ هستند. و این بدین معنی‌ست که هیچ ارتباطی بین متغیرهای پیشبینی کننده (x۱، x۲، x۳) و متغیر هدف (y) وجود ندارد.

$$H_0 : \beta_1 = 0 \quad and \quad \beta_2 = 0 \quad and \quad \beta_3 = 0 \tag{۱۹}$$

-فرض جایگزین، این را بیان می‌کند که حداقل یکی از ضرایب استفاده شده در مدل رگرسیون خطی، برابر ۰ نیست. و این بدین معنی‌ست که ارتباط significant ای بین متغیرهای پیشبینی کننده (x۱، x۲، x۳) و متغیر هدف (y) وجود دارد.

$$H_A : \beta_1 \neq 0 \quad or \quad \beta_2 \neq 0 \quad or \quad \beta_3 \neq 0 \tag{۲۰}$$

با توجه به جدول ANOVA، مقدار درجه آزادی رگرسیون برابر ۳ و درجه آزادی خطا برابر ۴۶ است. همچنین، مقدار آماره‌ی F برابر ۶.۵۷ است.

با توجه به این درجه‌های آزادی، و با در نظر گیری سطح significance برابر ۰.۰۵، مقدار بحرانی آماره‌ی F از نمودار مربوطه استخراج شد و برابر ۲.۸۱ است.
با توجه به این‌که مقدار آماره‌ی F مربوط به سوال، از مقدار بحرانی بدست آمده بیشتر است، شرط صفر رد می‌شود و نمی‌توان ادعا کرد که ضرایب استفاده شده در مدل رگرسیون خطی، همگی برابر ۰ هستند و هیچ ارتباطی بین متغیرهای پیشبینی کننده (x۱، x۲، x۳) و متغیر هدف (y) وجود ندارد.

۲-۴ زیربخش b

برای بررسی significant بودن هر یک از ضرایب مدل رگرسیون، باید برای هر یک از ضرایب، فرض صفر و فرض جایگزین را به شرح زیر تعریف کنیم:
-فرض صفر، این را بیان می‌کند که ضریب رگرسیون مدنظر، برابر ۰ است.

$$H_0 : \beta_i = 0 \quad (۲۱)$$

-فرض جایگزین، این را بیان می‌کند که ضریب رگرسیون مدنظر، برابر ۰ نیست.

$$H_A : \beta_i \neq 0 \quad (۲۲)$$

برای هر یک از ضرایب، با استفاده از فرمول زیر، آزمون T انجام می‌دهیم:

$$T_i = \frac{b_i - 0}{SE_i} \quad (۲۳)$$

-آماره‌ی T مربوط به beta۱:

$$T_1 = \frac{b_1 - 0}{SE_1} = \frac{90 - 0}{20} = 4.5 \quad (۲۴)$$

-آماره‌ی T مربوط به beta۲:

$$T_2 = \frac{b_2 - 0}{SE_2} = \frac{50 - 0}{15} = 3.33 \quad (۲۵)$$

-آماره‌ی T مربوط به beta۳:

$$T_3 = \frac{b_3 - 0}{SE_3} = \frac{10 - 0}{8} = 1.25 \quad (۲۶)$$

با در نظر گرفتن درجه آزادی برابر ۴۶ و سطح significance برابر ۰.۰۵، مقدار بحرانی آماره‌ی T برابر ۱.۶۷ بدست آمد.
با توجه به این‌که مقدار آماره‌ی T مربوط به دو ضریب beta۱ و beta۲ بزرگ‌تر از مقدار بحرانی مربوطه است، فرض صفر مربوط به این دو ضریب رد می‌شود. و نمی‌توان ادعا کرد که ضرایب beta۱ و beta۲ برابر ۰ هستند. این بدین معنی‌ست که نمی‌توان ادعا کرد که ارتباط significant ای بین هزینه‌ی تبلیغات تلویزیون و رادیو با درآمد حاصل از فروش شرکت وجود ندارد. اما مقدار آماره‌ی مربوط به ضریب beta۳ کوچک‌تر از مقدار بحرانی مربوطه است. و نمی‌توان فرض صفر مربوط به این ضریب را رد کرد. و می‌توان ادعا کرد که ضریب beta۳ برابر ۰ است. این بدین معنی‌ست که ارتباط significant ای بین هزینه‌ی تبلیغات روزنامه با درآمد حاصل از فروش شرکت وجود ندارد.

۵ داده‌های EEG (R)

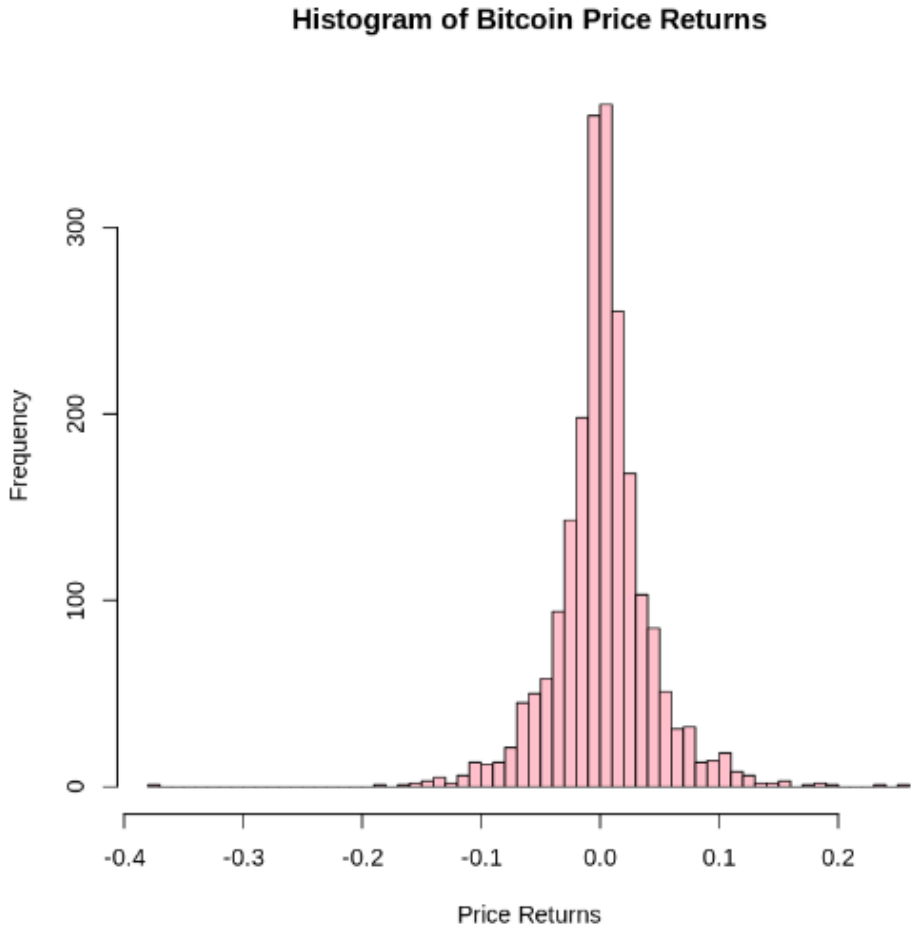
(از میان دو سوال ۵ و ۶، سوال ۶ پاسخ داده شده است.)

۶ داده‌های بیتکوین (R)

(از میان دو سوال ۵ و ۶، سوال ۶ پاسخ داده شده است.)

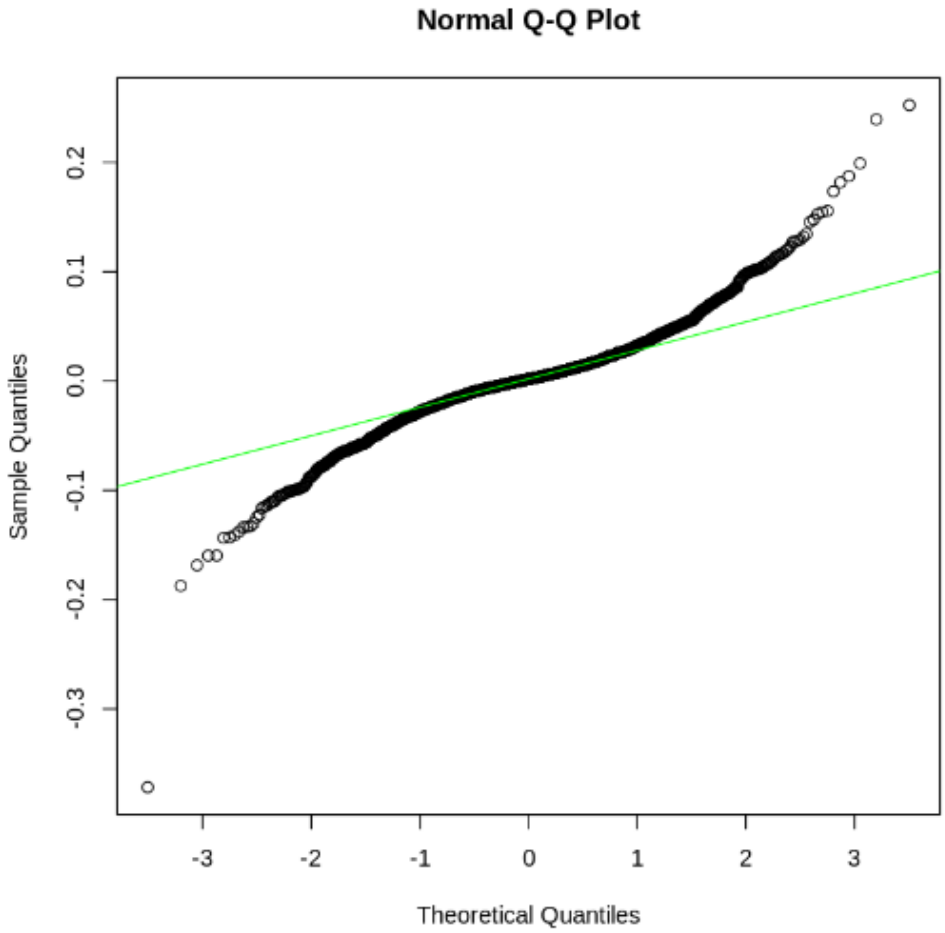
۶-۱ زیربخش ۱

نمودار histogram مربوط به price returns به فرم زیر است:



این نمودار، فرکانس توزیع فراوانی تغییرات درصدی قیمت بیتکوین را نشان می‌دهد. همانطور که در این نمودار نمایش داده شده است، این توزیع unimodal است و دارای توزیع زنگوله‌ای شکل تقریبا متقارن و در نتیجه دارای توزیع نرمال است. همچنین با توجه به این‌که peak نمودار حدوداً در ۰ رخ داده است، میانگین این توزیع برابر ۰ است. البته مشاهده می‌شود که همچنین دارای دم سنگین در سمت چپ است، که نشان‌دهنده‌ی تغییرات شدید منفی در قیمت بیتکوین است.

نمودار QQ-Plot مربوط به price returns به فرم زیر است:



با توجه به این‌که خط سبز رنگ رسم شده، مربوط به نمودار نرمال است، و نقاط نمودار تقریباً حول آن قرار دارند، از این نمودار هم این مسئله برداشت می‌شود، که توزیع مربوطه نرمال است. اگر هر دو طرف بالا یا هر طرف پایین بودند، توزیع دارای چولگی بود؛ که این طور نیست و در نتیجه توزیع مورد نظر تقریباً متقارن است.

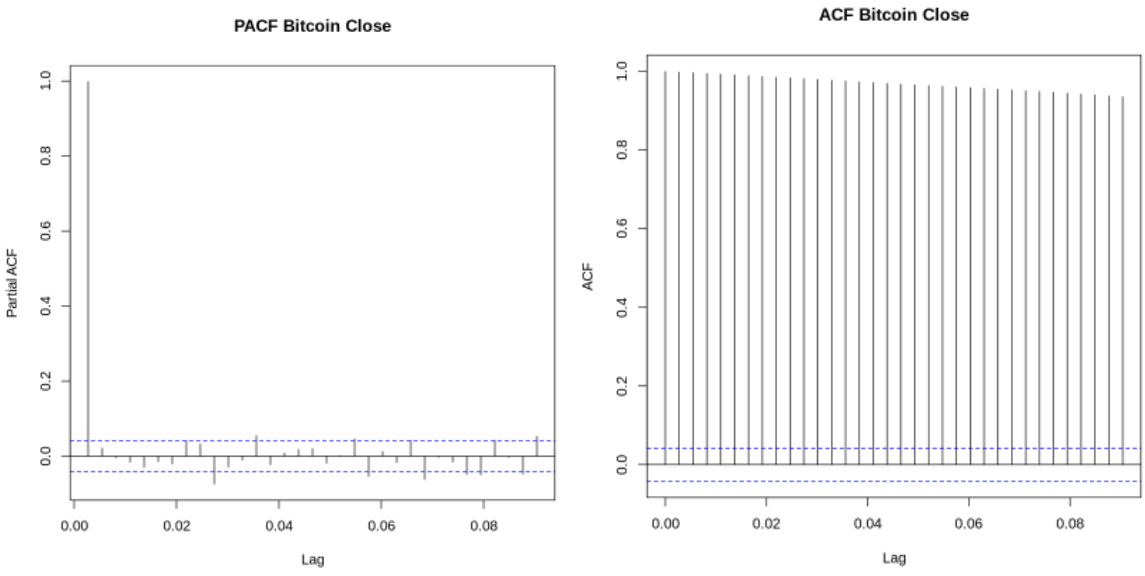
۲-۶ زیربخش ۲

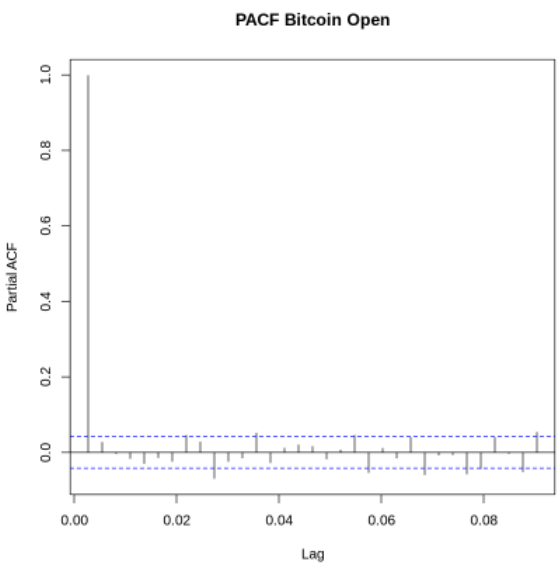
ACF همان تابع Autocorrelation است، که یک معیار آماری است و correlation بین یک سری زمانی و مقادیر lag آن را نشان می‌دهد. به عبارت دیگر، این را نشان می‌دهد که هر مشاهده در یک سری زمانی چقدر با مشاهدات قبلی خود مرتبط است. نمودار ACF نشان‌گر autocorrelation سری زمانی در lag های مختلف است.

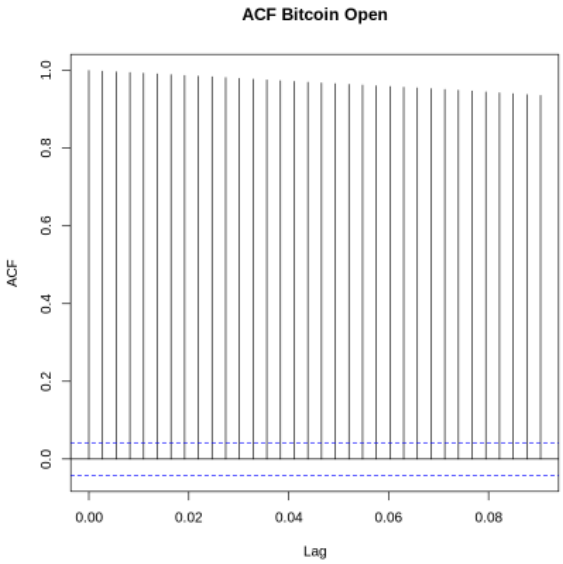
PACF همان تابع Autocorrelation جزئی است، که یک معیار آماری است و -corre lation بین یک سری زمانی و مقادیر lag آن را پس از حذف تاثیر lag های قبلی نشان می‌دهد. به عبارت دیگر، این را نشان می‌دهد که پس از در نظر گرفتن correlation با مشاهدات میانی، هر مشاهده در یک سری زمانی، چقدر با مشاهدات قبلی خود مرتبط است.

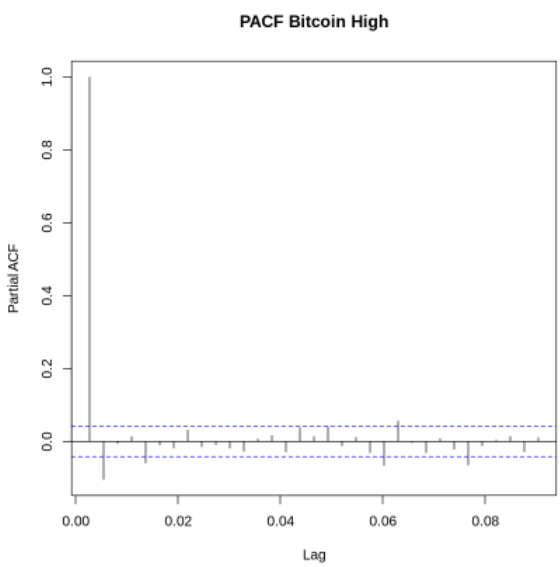
کاربرد این دو نمودار، شناسایی الگوها و correlation های موجود در داده‌هاست. با تحلیل این نمودارها، می‌توان دریافت کدام lag ها یا بازه‌های زمانی، با داده‌های سری زمانی correlation معنادار و قابل توجهی دارند و از این اطلاعات برای ساخت مدل‌های پیش‌بینی کننده استفاده کرد.

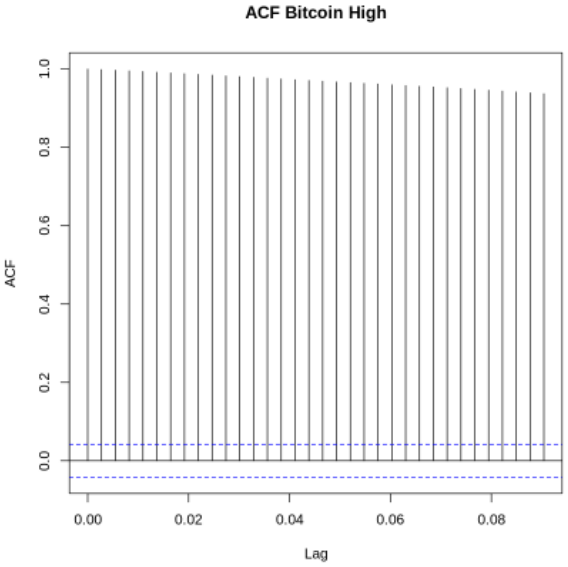
نمودارهای ACF و PACF مربوط به ۴ متغیر مربوط به قیمت در شکل‌های زیر قابل مشاهده است:

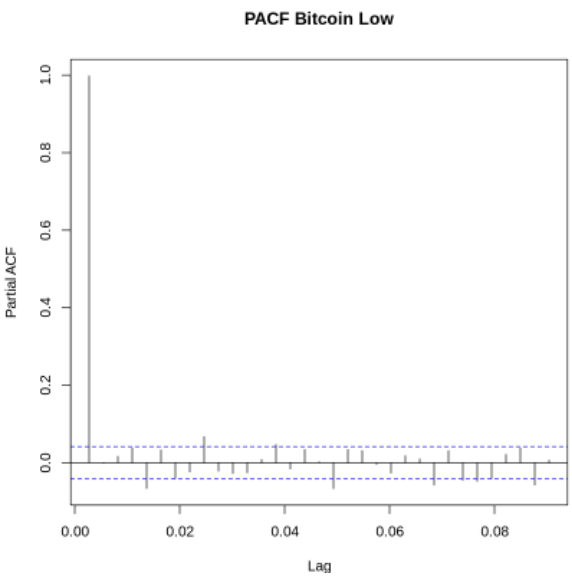


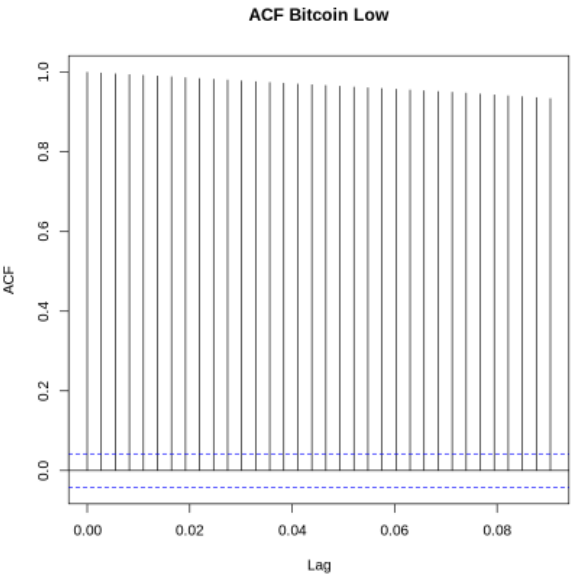

PACF Bitcoin Open
Partial ACF
Lag


ACF Bitcoin Open
ACF
Lag


PACF Bitcoin High
Partial ACF
Lag


ACF Bitcoin High
ACF
Lag


PACF Bitcoin Low
Partial ACF
Lag


ACF Bitcoin Low
ACF
Lag

در نمودار ACF مربوط به تمامی متغیرها، همه‌ی bar ها بالاتر از آستانه‌ی موردنظر برای significant بودن هستند. بدین معنی که میان این مقادیر و مقادیر دارای lag آن‌ها، رابطه‌ای وجود دارد. و با توجه به این‌که این bar ها همگی بالای آستانه‌ی مثبت هستند، همه‌ی این رابطه‌ها نیز، مثبت هستند. همچنین، طول این bar ها بلند است و این نشان‌گر این است که این روابط، همگی قوی هستند. همچنین، طول این bar ها به صورت نزولی است و این مسئله ممکن است نشان‌گر روند فصلی باشد.
در نمودار PACF مربوط به تمامی متغیرها، اکثر bar ها دارای طول کوتاهی هستند و از آستانه‌ی مورد نظر عبور نکرده‌اند. الگوی bar ها بدین شکل است که به صورت متناوب مثبت و منفی هستند. و فقط تعدادی از bar های منفی دارای طول بیشتر از آستانه هستند. این نشان‌گر این است که میان این مقادیر و تعدادی از مقادیر دارای lag آن‌ها، رابطه‌ی منفی وجود دارد. این تناوب مثبت و منفی bar ها، ممکن است نشان‌گر فرایند autoregressive باشد.
این رابطه‌هایی که در دو پاراگراف بالای از دو نمودار ACF و PACF برداشت شده است، به ترتیب، همان Autocorrelation و Partial Autocorrelation ایست که قبل از رسم نمودارها توضیح داده شد.

۳-۶ زیربخش ۳

stationary time series یک سری زمانی است که خصوصیات آماری آن مانند میانگین به مرور زمان تغییر نمی‌کنند. این بدان معناست که autocorrelation و autoco-variance آن در طول زمان ثابت باقی می‌ماند و به زمان وابسته نیست.
بخش I از مدل‌هایی مانند ARIMA مخفف کلمه "Integrated" است. این بخش تعداد دفعاتی را که باید تفاضل بین مشاهدات پشت سر هم را بگیریم تا سری زمانی پایدار شود، اندازه‌گیری می‌کند.
بنابراین، بخش I از مدل‌های ARIMA نقش حیاتی در رسیدن به پایداری بازی می‌کند. بدون پایدار کردن سری، ساخت مدل‌های سری زمانی کارآمد که می‌توانند الگوهای آینده در داده‌ها را درک کنند، دشوار است.

۴-۶ زیربخش ۴

در سری‌های زمانی، نمی‌توان به طور تصادفی نمونه‌هایی از میان داده‌ها انتخاب کرد و آن‌ها را به testSet یا trainSet اختصاص داد. زیرا استفاده از مقادیر آینده برای پیش‌بینی مقادیر گذشته معنایی ندارد. به عبارت ساده، ما می‌خواهیم در هنگام آموزش مدل خود، از نگاه به آینده جلوگیری کنیم. و با توجه به این‌که بین مشاهدات وابستگی زمانی وجود دارد، باید این رابطه را حین آزمایش حفظ کنیم.

- روشی که برای cross-validation مدل سری زمانی استفاده می‌شود، یک روش چرخشی است. این روش، با یک زیرمجموعه کوچک از داده‌ها به عنوان داده‌های آموزش شروع می‌کند و مدل را با استفاده از آن‌ها آموزش می‌دهد. برای داده‌های بعدی پیش‌بینی انجام می‌شود و سپس دقت پیش‌بینی برای نقاط داده‌ی پیش‌بینی شده بررسی می‌شود. سپس همان نقاط داده‌ای که پیش‌بینی شده‌اند، به عنوان بخشی از مجموعه داده‌های آموزش بعدی استفاده می‌شوند و برای نقاط داده‌ی بعدی پیش‌بینی انجام می‌دهد...

- دو تکنیک time series split cross-validation و blocked cross-validation نیز برای پیش‌بینی سری‌های زمانی کاربرد دارند.

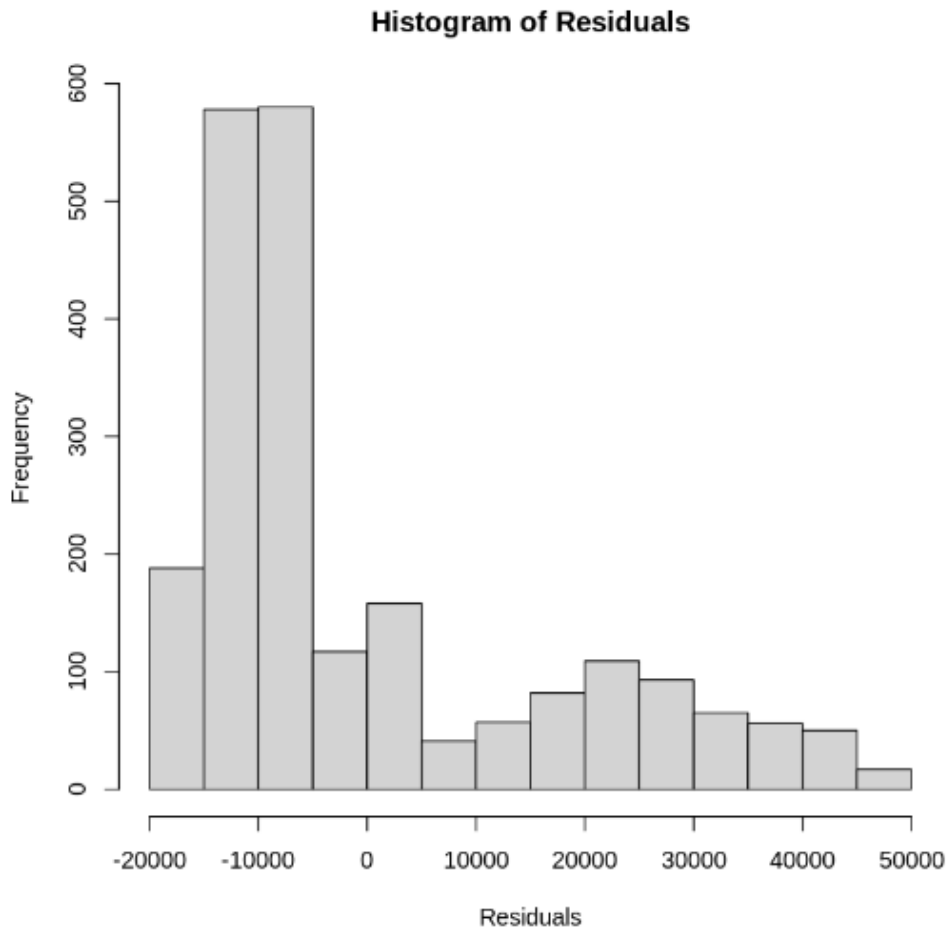
ایده‌ی تکنیک time series split cross-validation این است که trainSet را در هر تکرار به دو قسمت تقسیم کنیم، به شرطی که testSet همیشه از نظر زمانی جلوتر از trainSet باشد.

در blocked cross-validation، مدل، الگوهای آینده را برای پیش‌بینی مشاهده می‌کند و سعی می‌کند آن‌ها را به خاطر بسپارد. در دو موقعیت margin ایجاد می‌کند. اولین margin بین fold های trainSet و testSet است تا مدل از دو بار مشاهده شدن مقادیر lag ها، یک بار به عنوان regressor و بار دیگر به عنوان مقدار هدف، جلوگیری شود. margin دوم بین fold هایی است که در هر تکرار استفاده می‌شوند تا از به خاطر سپردن الگوها توسط مدل، از یک تکرار به تکرار بعدی جلوگیری شود.

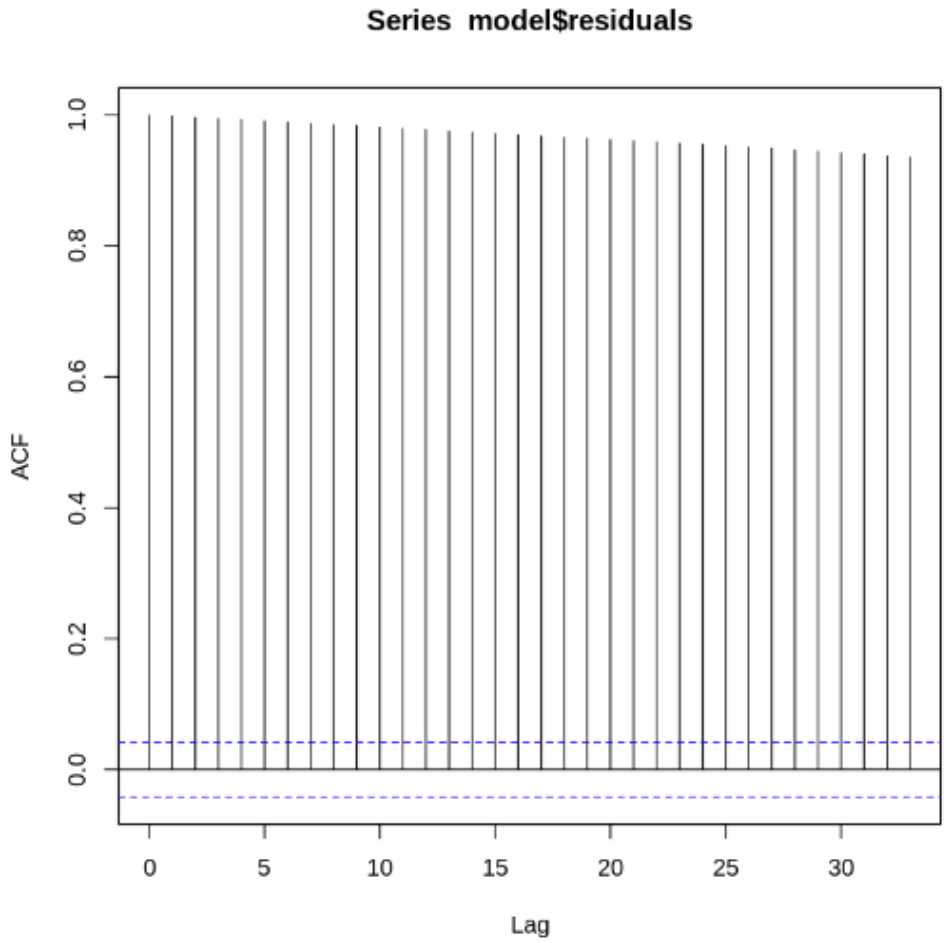
- برای nested cross-validation نیز، تکنیک Day Forward-Chaining و تکنیک Predict Second Half وجود دارد.

۵-۶ زیربخش ۵

نمودار histogram مربوط به residual ها به فرم زیر است:



برای بررسی وجود noise white نمودار ACF رسم شده است:



با توجه به نمودار histogram، توزیع به صورت نرمال نیست و میانگین آن ۰ نیست، در نتیجه white noise مشاهده نمی‌شود. همچنین، با توجه به این‌که در نمودار ACF طول bar ها از آستانه کم‌تر است و رابطه‌ی بین residual ها و lag های مختلف آن، قابل توجه نیست، white noise مشاهده نمی‌شود.