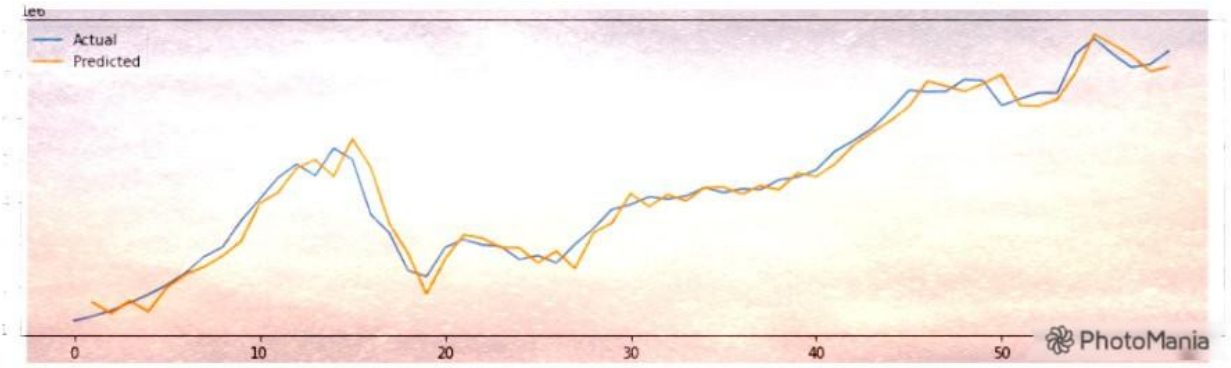
بسمه تعالی

فهرست مطالب

فصل ۱. مبانی نظری

تحلیل سری‌های زمانی

عبارت سری‌های زمانی به مجموعه‌داده‌هایی اطلاق می‌شود که دارای ترتیب زمانی هستند. دمای هوای متوسط روزانه‌ی یک منطقه در طول یک ماه، میزان بارندگی ماهانه و نرخ تورم سالانه، نمونه‌هایی از سری‌های زمانی شناخته‌شده هستند. تحلیل سری‌های زمانی از این جهت اهمیت دارد که معمولا در آن‌ها، مقادیر آینده‌ی سری، به روند تغییرات مقادیر گذشته‌ی آن وابسته است. پدیده‌های طبیعی و اقتصادی که به صورت سری‌های زمانی قابل توصیف هستند، دارای الگوهای نهفته‌ای هستند که با بررسی آن‌ها می‌توان مقادیر متغیر اصلی را در هر زمان از جمله در زمان‌های آینده، مدل و پیش‌بینی کرد. از این جهت حوزه‌ی تحلیل سری‌های زمانی از اهمیت بالایی برخوردار است.



تحلیل سری‌های زمانی به روش‌های مختلفی انجام می‌شود. از جمله‌ی این روش‌ها می‌توان به دو دسته‌ی کلی روش‌های آماری و روش‌های مبتنی بر یادگیری ماشین (از جمله مدل‌های LSTM) اشاره کرد. نکته‌ی قابل توجه آن است که روش‌های یادگیری ماشین، زمانی خوب عمل می‌کنند که داده‌های ورودی به اندازه‌ی کافی زیاد باشند. بنابراین این روش‌ها معمولا برای بررسی داده‌های کم حجم، از جمله پارامترهای کلان اقتصادی سالانه، خوب عمل نمی‌کنند. این در حالی است که در شرایط مشابه، مدل‌های آماری، تطابق بسیار خوبی با داده‌ها پیدا می‌کنند و پیش‌بینی‌های خوبی در اختیار می‌دهد. در طول این پروژه تمرکز بر روی روش‌های آماری (به طور خاص مدل‌های ARIMA) می‌باشد و این گفتار به آشنایی با مبانی نظری این مدل‌ها می‌پردازد.

مدل‌های ARIMA

مدل‌های ARIMA یکی از پرکاربردترین و جامع‌ترین مدل‌های آماری در تحلیل سری‌های زمانی هستند. در یک نگاه کلی، این دسته از مدل‌ها با در نظر گرفتن مقادیر پیشین و خطای تخمین یک متغیر در گام‌های زمانی گذشته، مقدار آن را در آینده پیش‌بینی می‌کنند. همانطور که از نام مدل‌های ARIMA بر می‌آید، این مدل‌ها از ترکیب سه بخش ساده‌تر AR، I و MA تشکیل شده‌اند که در ادامه به شرح آن‌ها پرداخته می‌شود.

مدل‌های AR (Autoregressive)

مدل‌های AR مقادیر آینده‌ی یک سری زمانی را با استفاده از مقادیر این متغیر در یک تا چند گام گذشته‌ی آن پیش‌بینی می‌کنند. در واقع در این مدل‌ها، مقدار متغیر به کمک ترکیب خطی‌ای از مقادیر پیشین خودش آموخته و پیش‌بینی می‌شود. رابطه‌ی تخمینی این مدل می‌تواند به صورت رابطه‌ی (1) تعریف شود.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} \qquad (1)$$

که در آن y_t مقدار متغیر مورد نظر در گام زمانی t، c یک ثابت و ϕ_t ضریب مقدار متغیر مورد نظر در گام زمانی t است.

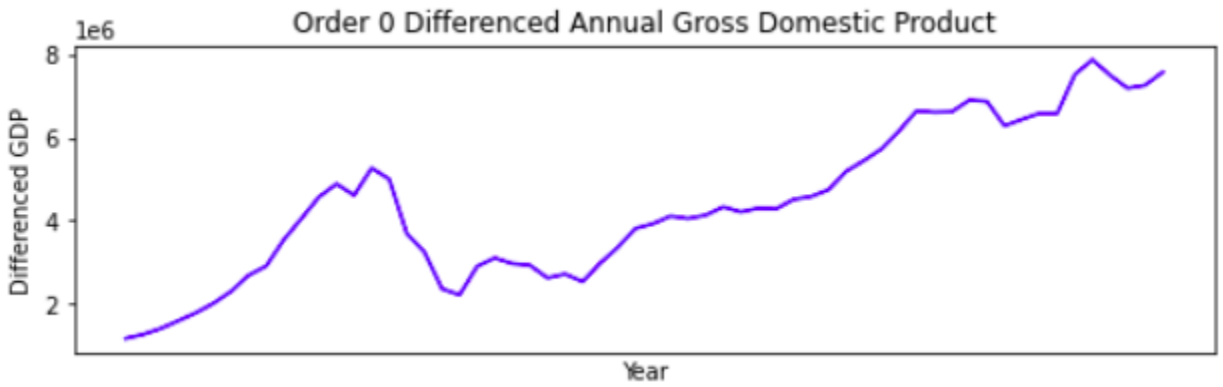
این رابطه به صورت رابطه‌ی (2) قابل بازنویسی است.

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} \qquad (2)$$

همانطور که از روابط (1) و(2) بر می‌آید، مدل‌های AR شامل یک پارامتر p هستند که تعداد گام‌های زمانی گذشته‌ی موثر در مقدار آینده‌ی متغیر را مشخص می‌کند. سایر ضرایب و ثوابت توسط مدل آموخته می‌شوند.

عبارت I (Integrated)

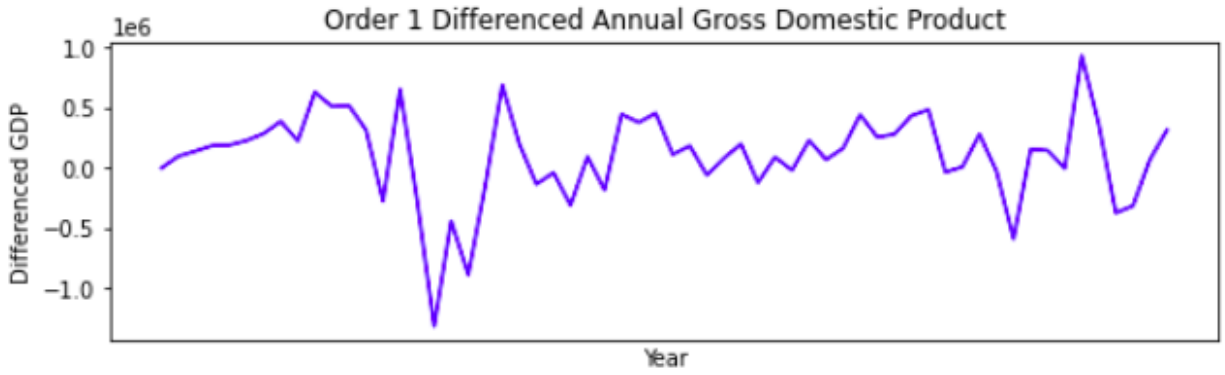
پیش‌فرض روابط و مفاهیم مورد استفاده در مدل‌های ARIMA، ایستایی (Stationarity) سری زمانی است. سری ایستا، دنباله‌ای از مقادیر است که میانگین و واریانس آن به زمان وابسته نباشد. به عبارتی سری زمانی باید عاری از هرگونه trend، seasonality و واریانس وابسته به زمان باشد. به عنوان مثال نمودار GDP سالانه که در تصویر (1) آمده‌است، دارای یک trend صعودی مشخص است که باعث ناایستایی آن می‌شود.



تصویر ۱. سری زمانی شاخص تولید ناخالص داخلی دارای روند صعودی بوده و ناایستا است.

در صورتی که سری زمانی ناایستا باشد، پیش‌فرض‌های مدل نقض می‌شوند و ممکن است نتایج نامناسبی را به‌دست بدهند. یکی از علل نیاز مدل‌های سری زمانی به ایستایی داده‌ها، پدیده‌ای به نام **spurious regression** است. سری‌های زمانی‌ای که روند یکسانی دارند، به عنوان مثال هر دو صعودی هستند، نسبت به هم **correlation** نشان می‌دهند. در حالی که ممکن است هیچ تاثیر علت و معلولی‌ای در رابطه با یکدیگر نداشته‌باشند. به همین ترتیب سری‌های زمانی غیر ایستا، می‌توانند **autocorrelation** گمراه‌کننده‌ای داشته باشند. مشکل دیگری که در سری‌های زمانی ناایستا وجود دارد، دقت نسبی مدل در زمان‌های مختلف است. به این ترتیب مدل‌های سری‌های زمانی نیازمند ایستایی میانگین، واریانس و همبستگی‌های ناوابسته به زمان هستند.

یکی از روش‌های از بین بردن ناایستایی در سری‌های زمانی، استفاده از **Difference Transform** است. در این روش با یک تا چند مرتبه کم کردن داده‌ها از مقادیر پیشین خود، سری زمانی می‌تواند به ایستایی مطلوبی برسد. به عنوان مثال سری زمانی شاخص تولید ناخالص داخلی که در تصویر (1) آمد، با یک مرتبه کم کردن داده‌ها از یک گام زمانی پیشین خود، به صورت تصویر (2) در آمده و به ایستایی می‌رسد.



تصویر ۲. سری زمانی شاخص تولید ناخالص داخلی با differencing مرتبه‌ی اول به یک سری ایستا تبدیل می‌شود.

مدل سپس با سری زمانی ایستا تغذیه می‌شود تا بتواند الگوهای واقعی و موثر موجود در آن را یاد بگیرد. در آخر لازم است یک inverse difference transform بر روی پیش‌بینی‌های مدل اعمال شود تا مقادیر به order واقعی و پیش از differencing خود بازگردند.

با توضیحی که گذشت، می‌توان گفت عبارت I در مدل‌های ARIMA نشان‌دهنده‌ی تعداد دفعات Differencing بر روی سری زمانی برای رسیدن به ایستایی مطلوب را نشان می‌دهد. در واقع مدل‌های ARIMA، عملیات difference trainform و inverse difference transform را، که معمولا به صورت دو مرحله‌ی preprocessing و post-processing انجام می‌شود، در خود مدل تعبیه کرده‌اند.

مدل‌های MA (Moving Average)

مدل‌های MA مقادیر آینده‌ی یک سری زمانی را با استفاده از مقدار خطاهای تخمین آن در یک تا چند گام گذشته پیش‌بینی می‌کنند. در واقع در این مدل‌ها، مقدار متغیر به کمک ترکیب خطی‌ای از مقادیر اخیر خطای تخمین خودش آموخته و پیش‌بینی می‌شود. رابطه‌ی تخمینی این مدل می‌تواند به صورت رابطه‌ی (3) تعریف شود.

$$y_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} \qquad (3)$$

که در آن y_t مقدار متغیر مورد نظر در گام زمانی t، c یک ثابت، θ_t ضریب مقدار متغیر مورد نظر در گام زمانی t و ϵ_t مقدار خطا یا به عبارتی اختلاف مقدار تخمینی برای متغیر در زمان t با مقدار واقعی آن در این زمان است.

این رابطه به صورت رابطه‌ی (4) قابل بازنویسی است.

$$y_t = c + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \qquad (4)$$

همانطور که از روابط (3) و(4) بر می‌آید، مدل‌های MA شامل یک پارامتر q هستند که تعداد گام‌های زمانی گذشته‌ی موثر در مقدار آینده‌ی متغیر را مشخص می‌کند. سایر ضرایب و ثوابت توسط مدل آموخته می‌شوند.

مدل ترکیبی ARMA

مدل ARMA همانگونه که از نام آن بر می‌آید، از ترکیب مدل‌های AR و MA ساخته می‌شود و رابطه‌ی تخمینی آن به صورت رابطه‌ی (5) تعریف می‌شود.

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \qquad (5)$$

بدیهی است که این مدل شامل دو پارامتر p و q خواهد بود که شرح آن در مدل‌های AR و MA آمد.

مدل ترکیبی ARIMA

تعریف این مدل مشابه مدل ARMA است با این تفاوت که به جای مقادیر y_t با مقادیر تبدیل‌یافته‌ی آن یعنی $y_t^{[d]}$ سروکار دارد. در این نماد، d نشان‌دهنده‌ی درجه‌ی differencing بوده و $y_t^{[d]} = y_t^{[d-1]} - y_{t-1}^{[d-1]}$ است. برای نمایش عملیات differencing می‌توان از عملگر Δ^d استفاده کرد. به گونه‌ای که $\Delta^d x = x^{[d]}$. به این ترتیب رابطه‌ی تخمینی مدل ARIMA به صورت رابطه‌ی (6) قابل نمایش است.

$$\Delta^d y_t = c + \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \sum_{i=1}^{q} \theta_i \Delta^d \epsilon_{t-i} \qquad (6)$$

مدل‌های ARIMAX (Future Work)

در قسمت‌های قبل، مدل‌های **ARIMAX** به عنوان یکی از شناخته‌شده‌ترین مدل‌های سری زمانی معرفی شدند. این مدل‌ها برای آنالیز یک سری زمانی بر اساس مقادیر و خطاهای تخمینی پیشین خود آن سری به کار می‌آیند. در این بخش مدل‌های **ARIMAX** معرفی می‌شوند. این مدل‌ها حالت عمومی‌تری از مدل‌های **ARIMA** شامل متغیرهای خارجی موثر بر متغیر مورد علاقه هستند. در واقع مدل‌های **ARIMAX** یک ترکیب خطی از مقادیر متغیرهای خارجی در هر گام زمانی را نیز در تخمین و پیش‌بینی متغیر مورد بررسی، دخالت می‌دهند.

رابطه‌ی تخمینی مدل ARIMAX

رابطه‌ی این مدل از گسترش رابطه‌ی مدل **ARIMA** با ترکیب خطی متغیرهای خارجی به‌دست می‌آید. به طوری که اگر **m** متغیر خارجی را در مقادیر متغیر مورد بررسی دخیل بدانیم، رابطه‌ی تخمینی مدل به صورت رابطه‌ی **(7)** خواهد بود.

$$\Delta^d y_t = c + \sum_{i=1}^{p} \phi_i \Delta^d y_{t-i} + \sum_{i=1}^{q} \theta_i \Delta^d \epsilon_{t-i} + \sum_{i=1}^{m} \beta_i x_{ti} \qquad (7)$$

که در آن x_{ti} مقدار متغیر خارجی i ام در گام زمانی t و β_i ضریب متغیر خارجی i ام در ترکیب خطی متغیرهای خارجی است.

توجه شود که مقادیر متغیرهای خارجی در این رابطه به صورت **autoregressive** دخیل نیست. بلکه مقدار مورد پیش‌بینی **y** در گام زمانی **t** تنها از مقدار متغیرهای خارجی در همان گام زمانی تاثیر مستقیم می‌پذیرد.

پیاده‌سازی

آنالیز سری‌های زمانی به کمک مدل‌های **ARIMA(X)**، شامل انتخاب و تعریف پارامترهای مدل، **fit** کردن مدل بر روی داده‌های **train**، ارزیابی مدل و انجام پیش‌بینی به کمک مدل تاییدشده است. در این پروژه، الگوریتم **Box-Jenkins** برای پیاده‌سازی و ارزیابی مدل‌ها مبنا قرار داده می‌شود. **Box-Jenkins** که توسط سازندگان مدل **ARIMA** پیشنهاد داده شده‌است، یک روش قاعده‌مند و مورد اطمینان برای تعریف و ارزیابی این مدل‌ها ارائه می‌دهد. شرح مراحل و متدهای پیشنهادی این الگوریتم در ادامه آمده‌است.

مرحله‌ی ۱ - Model Identification & Model Selection

این مرحله شامل کسب اطمینان از Stationary بودن داده‌ها و انتخاب پارامترهای مدل است.

تنظیم Stationarity داده‌ها

شرح لزوم stationarity در بخش تعاریف گذشت. در این بخش افزودن دو نکته خالی از لطف نیست. اول، محک تشخیص استایی داده‌ها و دوم، روش انتخاب پارامتر d و یا همان درجه‌ی differencing.

تست‌ها و روش‌های مختلفی برای تشخیص ایستایی سری‌های زمانی وجود دارد. یکی از رایج‌ترین این تست‌ها که در این پژوهش نیز مورد استفاده قرار گرفته است، تست Augmented-Dickey Fuller یا به اختصار ADF می‌باشد. Null hypothesis این تست، Non-Stationary بودن داده‌ها است. بنابراین در صورتی که اعمال تست بر روی داده‌ها به p-value ای کمتر از ۰.۰۵ منجر بشود، با اطمینان بیش از ۹۵ درصد می‌توان ادعا کرد که داده‌ها Stationary هستند. به این ترتیب می‌توان مقادیر مختلف پارامتر d را به کمک این تست آزمایش کرد و کوچک‌ترین درجه‌ی differencing که باعث پاس شدن تست ایستایی می‌شود را برگزید.

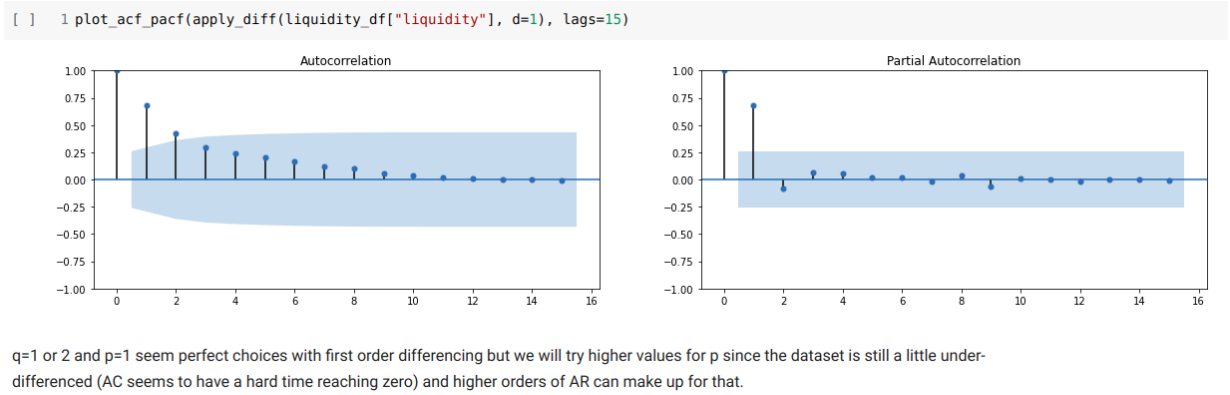
اما نکته‌ی قابل توجه آن است که این مقدار پارامتر d در عمل لزوما به این شکل انتخاب نمی‌شود. چرا که در رابطه با ایستا سازی داده‌ها، باید علاوه بر پدیده‌ی Under-Differencing (نرسیدن داده‌ها به ایستایی لازم)، به پدیده‌ی Over-Differencing نیز توجه داشت. Over-Differencing زمانی رخ می‌دهد که با اعمال Difference Transform های بیش‌تر از حد مناسب، اگرچه دیتاست به حالت ایستایی برسد، ویژگی‌ها برجسته‌ی سری‌زمانی محو می‌شود و مدل نمی‌تواند یادگیری هوشمندانه‌ای بر روی داده‌ها انجام دهد و از الگوهای موجود در آن بهره‌ی لازم را ببرد.

برخی محک‌ها وجود دارند که در تشخیص Under-Differencing و Over-Differencing کمک‌کننده هستند. ۱. اگر سری‌زمانی برای تعداد زیادی گام زمانی Autocorrelation مثبتی را نشان دهد، نیاز به differencing بیش‌تری وجود دارد. ۲. اگر Autocorrelation برای یک گام زمانی، صفر یا منفی باشد و یا نمودار Autocorrelation مقادیر کوچک و بی‌نظمی را نشان دهد، differencing بیش‌تری نیاز نیست. ۳. در صورتی که Autocorrelation برای یک گام زمانی، مقدار منفی بزرگی باشد (منفی ۰.۵ یا کمتر) احتمالا سری زمانی دچار Over-differencing شده‌است. ۴. در دو حالت differencing با شرایط نسبتا مشابه، اولویت با حالتی است که داده‌ها در آن STD کمتری داشته‌باشند. در این پژوه درجه‌ی differencing با بهره‌گیری از این روش‌ها و روش Model Selection ای که در ادامه خواهد آمد، انتخاب می‌شوند.

انتخاب پارامترها

یک انتخاب اولیه و مناسب برای پارامترهای مدل ARMA می‌تواند به کمک نمودارهای ACF و PACF یا به عبارتی Partial Autocorrelation Function انجام شود. این نمودارها با یک بازه‌ی اطمینان، گام‌هایی را نشان می‌دهند که در آن داده‌ها به احتمال بالایی دارای یک AC/PAC قوی هستند. این یعنی همان پارامترهای موثر AR و MA. چرا که این مدل‌ها از مقادیر چند گام قبل برای پیش‌بینی گام بعدی استفاده می‌کنند و به طور ایده‌آل این رابطه را بر روی تعداد گامی اجرا می‌کنند که همبستگی قوی‌ای را نشان می‌دهند.

به این ترتیب برای تعیین پارامتر AR تعداد گام‌های زمانی‌ای انتخاب می‌شوند که در آن‌ها بزرگی PAC از حد اطمینان بیش‌تر و برای تعیین پارامتر MA تعداد گام‌های زمانی‌ای انتخاب می‌شوند که بزرگی AC آن‌ها بزرگ‌تر از بازه‌ی اطمینان باشد. تصویر ۳ یک نمونه انتخاب پارامتر بر اساس این نمودارها برای سری زمانی نقدینگی را نشان می‌دهد.



تصویر ۳. انتخاب پارامترهای p و q با استفاده از ACF و PACF سری زمانی نقدینگی که با یک مرتبه differencing ایستا شده است.

روش فوق‌الذکر برای انتخاب پارامترهای مدل برای انجام بهترین انتخاب کافی نیست. زیرا این روش ممکن است چند ترکیب مختلف از پارامترهای p و q را به‌دست دهد. و یا ممکن است نمودارهای PACF و ACF بی‌نظم بوده یا همبستگی قوی و مشخصی را نشان ندهند و انتخاب پارامترها را سخت کنند. بنابراین معرفی روش دیگری برای انتخاب مدل الزامی است. این روش با استفاده از Information Criteria، مدل‌های مختلف را مقایسه و بهترین مدل را معرفی می‌کند.

دو مورد از معیارهای Information Criteria ای که می‌توانند در این روش مورد استفاده قرار بگیرند، AIC و BIC هستند که در این پژوهش از مورد اول برای model selection استفاده شده‌است. AIC به صورت رابطه‌ی (8) تعریف می‌شود.

$$AIC = -2log(L) + 2(p + q + k + 1) \qquad (8)$$

که در آن L نشان‌دهنده‌ی likelihood داده‌ها بوده و $k = 1$ است اگر $c \neq 0$ و $k = 0$ است اگر $c = 0$ باشد.

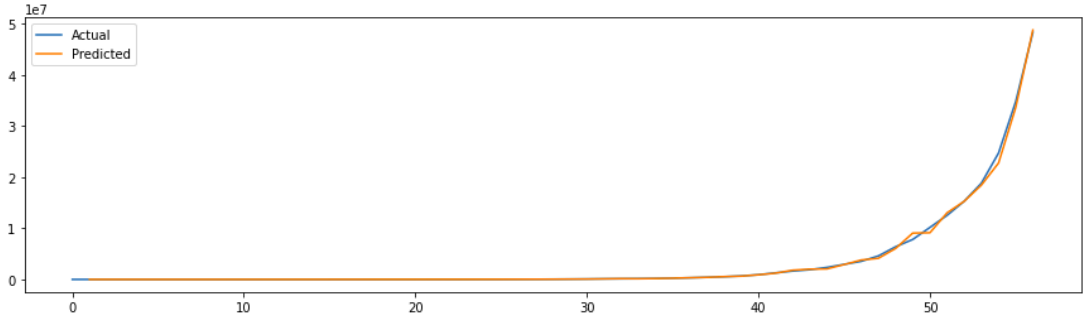
در این رابطه از تابع likelihood استفاده شده‌است. این تابع برابر است با احتمال آنکه به ازای پارامترهای یک مدل، نتایج مشاهده شده حاصل شود. معیار AIC جمله‌ی log likelihood را با علامت منفی شامل می‌شود. بنابراین هر چه یک مدل، بهتر و انتخاب پارامترهای آن برای داده‌ها مناسب‌تر باشد، AIC آن مدل عدد کوچک‌تری خواهد بود. علاوه بر این، AIC جمله‌ی جمع پارامترها را با علامت مثبت دارد. در نتیجه هر چه مقدار پارامترها بزرگ‌تر و به عبارتی مدل، پیچیده‌تر باشد، AIC آن مدل عدد بزرگ‌تری می‌شود. واضح است که در انتخاب بین دو مدل، مدل با AIC کوچک‌تر مطلوب می‌باشد. به این ترتیب معیار AIC، خوب بودن fit مدل بر داده‌ها و ساده بودن مدل را همزمان در رابطه‌ی خود اعمال می‌کند و به این ترتیب از overfitting به واسطه‌ی پارامترهای بزرگ جلوگیری می‌کند.

یکی از روش‌های انتخاب مدل بر اساس معیار AIC، استفاده از کتاب‌خانه‌ی pmdarima است. این کتاب‌خانه ماژولی به نام Auto ARIMA ارائه می‌دهد که با انجام یک Grid Search بر روی ترکیب‌های مختلفی از پارامترها و محاسبه‌ی AIC برای مدل‌های حاصل، مدل با کمترین AIC را انتخاب می‌کند. یک نمونه از model selection توسط این utility را در تصویر ۴ مشاهده می‌کنید.

```
1 apply_auto_arima_model_selection(liquidity_df["liquidity"], stepwise=False)
```

```
ARIMA(0,4,0)(0,0,0)[0]             : AIC=1570.348, Time=0.03 sec
ARIMA(0,4,1)(0,0,0)[0]             : AIC=1535.356, Time=0.13 sec
ARIMA(0,4,2)(0,0,0)[0]             : AIC=1536.926, Time=0.16 sec
ARIMA(0,4,3)(0,0,0)[0]             : AIC=1538.898, Time=0.39 sec
ARIMA(0,4,4)(0,0,0)[0]             : AIC=1537.289, Time=0.51 sec
ARIMA(0,4,5)(0,0,0)[0]             : AIC=1546.247, Time=0.49 sec
ARIMA(1,4,0)(0,0,0)[0]             : AIC=1547.628, Time=0.04 sec
ARIMA(1,4,1)(0,0,0)[0]             : AIC=1534.346, Time=0.13 sec
ARIMA(1,4,2)(0,0,0)[0]             : AIC=1536.089, Time=0.27 sec
ARIMA(1,4,3)(0,0,0)[0]             : AIC=1538.965, Time=0.35 sec
ARIMA(1,4,4)(0,0,0)[0]             : AIC=1538.410, Time=0.87 sec
ARIMA(2,4,0)(0,0,0)[0]             : AIC=1538.395, Time=0.04 sec
ARIMA(2,4,1)(0,0,0)[0]             : AIC=1536.161, Time=0.30 sec
ARIMA(2,4,2)(0,0,0)[0]             : AIC=inf, Time=1.14 sec
ARIMA(2,4,3)(0,0,0)[0]             : AIC=1534.054, Time=1.06 sec
ARIMA(3,4,0)(0,0,0)[0]             : AIC=1538.788, Time=0.14 sec
ARIMA(3,4,1)(0,0,0)[0]             : AIC=1538.152, Time=0.45 sec
ARIMA(3,4,2)(0,0,0)[0]             : AIC=inf, Time=1.31 sec
ARIMA(4,4,0)(0,0,0)[0]             : AIC=1535.981, Time=0.05 sec
ARIMA(4,4,1)(0,0,0)[0]             : AIC=1538.430, Time=0.27 sec
ARIMA(5,4,0)(0,0,0)[0]             : AIC=inf, Time=0.95 sec

Best model:  ARIMA(2,4,3)(0,0,0)[0]
Total fit time: 9.214 seconds
```



تصویر ۴. Model Selection به کمک Auto ARIMA

در scope مدنظر این پروژه انتخاب مدل به کمک کتابخانه‌ی pmdarima انجام می‌شود و Manual Parameter Tuning به عنوان Future Work در بک‌لاگ پروژه قرار می‌گیرد.

مرحله‌ی ۲ - Parameter Estimation

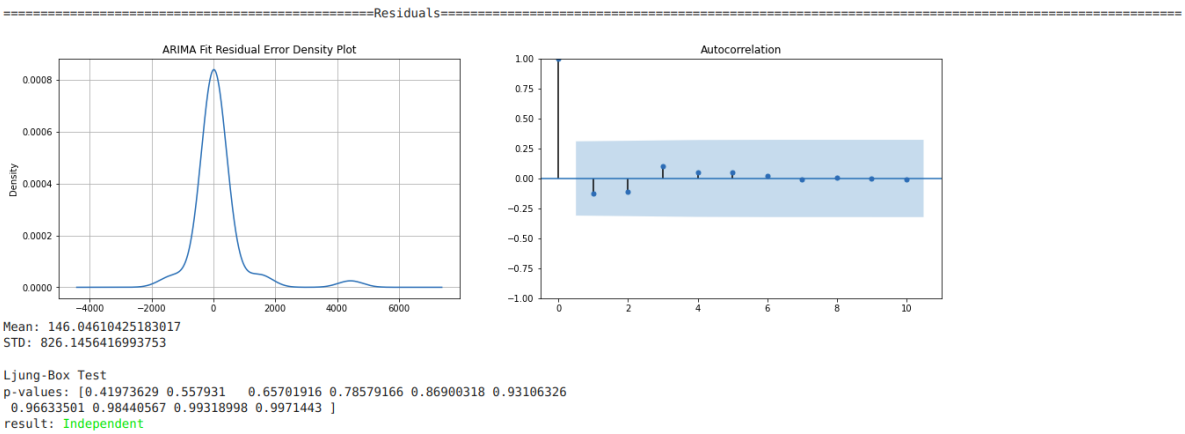
این مرحله که شامل تخمین مقادیر مناسب برای پارامترهای داخلی مدل از جمله θ_i ها و ϕ_i ها می‌باشد، به کمک تابع fit مدل‌ها از کتابخانه‌های آماده انجام می‌شود و جزئیات پیاده‌سازی آن در این گفتار نمی‌آید.

مرحله‌ی ۳ - Model Checking

پس از انتخاب و fit کردن مدل‌ها، نیاز است که مدل ارزیابی شود. باید بتوان از عملکرد مدل در پیش‌بینی‌های Out-of-Sample به نحوی اطمینان حاصل کرد. چراکه عملکرد خوب مدل بر روی داده‌های train لزوما به معنی دقیق بودن آن در گام‌های زمانی آینده نیست و ممکن است ناشی از overfitting مدل باشد. همچنین دقت داریم که روش‌های مرسوم ارزیابی مدل به کمک train-test split در سری‌های زمانی با مشکلاتی مواجه هستند. اولین مشکل که در تحلیل سری‌های زمانی اقتصادی نیز بسیار متداول است، تعداد بسیار محدود داده‌ها است و این امر مانع از جدا سازی بخشی از داده‌ها از مجموعه‌ی train می‌شود. مشکل دیگر از آنجا نشأت می‌گیرد که مهم‌ترین اطلاعات مورد ارزیابی در یک سری زمانی، اطلاعات گام‌های زمانی پایانی هستند. چرا که مقدار یک متغیر در یک گام زمانی به احتمال زیاد به الگوی گام‌های زمانی نزدیک به آن وابستگی و شباهت بیشتری دارد. بنابراین جدا کردن گام‌های زمانی پایانی به عنوان مجموعه‌ی test، مهم‌ترین داده‌ها برای تحلیل مقدار متغیر مورد علاقه در گام‌های زمانی آینده را از بین می‌برد و در نتیجه گزینه‌ی مناسبی نیست.

الگوریتم Box-Jenkins روشی را برای ارزیابی مدل‌های ARIMA(X) معرفی کرده که مبتنی بر خطاهای مدل در پیش‌بینی‌های In-Sample است. این روش عنوان می‌دارد که یک مدل تنها در صورتی پذیرفته و قابل اتکا است که Residual ها یا خطاهای آن از هم مستقل باشند و Autocorrelation نداشته باشند. به این منظور در بخش ارزیابی مدل‌ها از دو تست Ljung-Box و بررسی نمودارهای Residual ها استفاده شده‌است.

تست Ljung-Box یک تست آماری با Null Hypothesis مستقل بودن داده‌ها می‌باشد. بنابراین یک p-value بزرگ‌تر از ۰.۰۵ با اطمینان ۹۵ درصد از استقلال داده‌ها خبر می‌دهد. این تست بر روی یک تعداد گام مشخص از داده‌ها انجام می‌شود. به عنوان مثال، تست Ljung-Box مرتبه‌ی ۱ نشان دهنده‌ی وجود/عدم وجود Autocorrelation در داده‌ها با یک گام lag می‌باشد. در این پژوهش تست بر روی ۱۰ تعداد گام انجام شده‌است و در صورتی که تمام تعداد گام‌ها تست را پاس کنند، Residual ها مستقل اعلام می‌شوند. یک نمونه از اعمال این تست بر روی خطاهای مدل را در تصویر ۵ مشاهده می‌کنید.



تصویر ۵. نمودارها و تست بررسی Independence خطاهای مدل.

این تصویر همچنین نبود Autocorrelation قوی بین Residual ها را به کمک نمودار ACF تایید می‌کند.

فصل ۲. معرفی محصول

فصل ۳. فرایندهای اسکرام

فصل ۴. جزئیات فنی