Ahmet Çeliker
2020402126

IE 360 HOMEWORK1 REPORT

Research Model 1

Research Question

"Does the volume of credit sales (Kredi Hacmi) exhibit a relationship with changes in the interest rate on deposits (Mevduat Faiz Oranı), incorporating both trend and seasonality effects?"

This question aligns with my analysis, which includes exploring the relationship between credit sales and deposit interest rates while considering trend and seasonality components. It allows to investigate how fluctuations in deposit interest rates, along with other temporal patterns, may impact credit sales volume.

Data Acquisition and Preparation: I kickstart my exploration by sourcing data from an Excel file, leveraging the **read.xlsx** function to seamlessly import it into my R environment. This dataset, encapsulated within a data frame named **kredi**, houses two pivotal variables: credit volume (**KrediHacmi**) and deposit interest rates (**MevduatFaizOrani**). Subsequently, I transform the data frame into a **data.table** object for streamlined manipulation.

Visualizing Insights: My quest for understanding the dynamics between credit volume and deposit interest rates unfolds through captivating visualizations. I first wield **ggplot** to craft a scatter plot, vividly portraying the relationship between deposit interest rates and credit volume. This graphical representation offers a tangible glimpse into potential correlations.

To gain a holistic understanding of inter-variable relationships, I harness **ggcorrplot** to render a correlation matrix plot. This visual gem unveils the intricate web of correlations between all variables in my dataset, shedding light on potential dependencies.

Delving deeper into temporal dynamics, I construct time series plots, showcasing the evolution of credit volume and deposit interest rates over time. These plots not only highlight trends but also unveil patterns, guiding my quest for insights.

Moreover, my exploration extends to seasonal patterns, as evidenced by the creation of seasonal plots. These visually captivating representations uncover recurring trends in credit volume, offering valuable insights into seasonality effects.

Unearthing Correlations: To quantify relationships, I conduct correlation analyses using the **cor** function. This statistical inquiry unveils the degree of association between deposit interest rates and credit volume, allowing me to discern meaningful patterns.

Insights and Reflections: My journey through data manipulation and visualization has unearthed intriguing insights into the relationship between credit volume and deposit interest rates. From visually striking plots to statistically significant correlations, my analysis offers a comprehensive understanding of this dynamic interplay.

Model Building: I begin by fitting a basic linear regression model (**l_fit**) with credit volume (**KrediHacmi**) as the dependent variable and deposit interest rates (**MevduatFaizOrani**) as the independent variable. This model provides a baseline understanding of the relationship between the two variables.

Subsequently, I explore more complex models to account for additional factors influencing credit volume. I incorporate a trend variable (**trnd**) to capture any underlying trends in the data and fit a regression model (**lm_trend**) to assess its impact on credit volume.

Model Evaluation: To evaluate the validity of the regression models, I conduct various statistical tests and diagnostic checks. These include:

- Residual Analysis: I inspect the residuals of each model to ensure that they meet the assumptions of linear regression, such as homoscedasticity and normality.
- Significance Testing: I analyze the significance of coefficients in each model to determine the strength and direction of relationships between variables.
- Model Comparison: I compare the goodness-of-fit statistics of different models to identify the most suitable model for predicting credit volume.

Building a Linear Regression Model:

I built a linear regression model using lm. The model tries to explain the changes in "KrediHacmi" (loan volume) based on the changes in "MevduatFaizOrani" (deposit interest rate). I stored the model in a variable named l_fit.

I then used summary to get a detailed report on the model's performance. This report includes information about the coefficients of the model, their standard errors, p-values, and the overall R-squared value, which tells me how well the model fits the data.

I used the plot function to generate a diagnostic plot of the model. This plot typically shows the residuals (errors) of the model versus the fitted values. It helps me diagnose potential issues with the model.

I added a smooth line (geom_smooth) to the scatter plot I created earlier. This line represents the fitted model based on linear regression (method='lm'). This helps visualize how well the model captures the trend in the data.

I decided to improve the model by including a time trend. I added a variable named trnd to the data, where each data point gets a number representing its order (likely corresponding to time). Then, I refitted the model with this new trend term included as an additional predictor.

I summarized the new model with the trend term using summary and plotted the fitted line to see how the trend affects the relationship between the variables.

Checking Model Assumptions:

For linear regression models to be reliable, there are certain assumptions that need to be met. One assumption is that the variance of the errors (residuals) should be constant throughout the data. I performed a Breusch-Pagan test (bptest) to check for this assumption violation called heteroscedasticity.

Looking at the model with the trend (l_fit_trend), I created a new ggplot that includes the fitted line from the model.

I investigated the normality of the residuals from the model with trend (l_fit_trend$residuals) using checkresiduals. Normality is another assumption for linear regression.

I performed Augmented Dickey-Fuller (ADF) tests to see if the original KreditHacmi series and the residuals of the model with trend are stationary. Stationarity means the data doesn't exhibit trends or seasonal patterns over time, which is an important assumption for linear regression models.

Differencing and Log Transformation:

I took another look at the summary of the data (kredi) using summary and str to see if there's anything else I might be missing.

I created a differenced version of the data (kredi_diff) by subtracting consecutive values for "KrediHacmi" and "MevduatFaizOrani". This differencing technique is a common way to achieve stationarity in time series data.

I fit a new model (l_fit_diff_simple) using the differenced data, with only "MevduatFaizOrani" as the predictor to see see how the relationship between the differenced variables looks like. I visualized this with a scatter plot and a smooth line (geom_smooth) to see if differencing helped.

I noticed that the "KrediHacmi" variable might not be normally distributed or have constant variance. To address this, I applied a log transformation (log(x+1)) to "KrediHacmi". This helps with potential non-normality or heteroscedasticity issues.

I then refitted a new model (l_fit_log) with the log-transformed "KrediHacmi" as the dependent variable. I included "MevduatFaizOrani" and the trend term (trnd) as predictors in this model.

Ahmet Çeliker
2020402126

I examined the new model with the log transformation using summary to see how well it performs. I also checked the residuals of this model for normality using checkresiduals.

Finally, I performed ADF tests again on the log-transformed KreditHacmi series and the residuals of the new model to see if stationarity was achieved.

Weighted Least Squares (WLS):

The previous model might still have an issue with unequal variance of the errors (heteroscedasticity). To address this, I calculated weights based on the absolute values of the residuals from the previous model (l_fit_log).

I then employed a technique called Weighted Least Squares (WLS) by fitting a new model (l_fit_wls) using these weights. WLS gives more weight to observations with smaller residuals and less weight to those with larger residuals, potentially leading to a more accurate model in case of heteroscedasticity.

I'll explore the results of the WLS model further to see if it improves upon the previous ones.

Absolutely, here's the explanation of the code snippet in your first-person perspective:

Capturing Seasonality with Fourier Features:

I decided to account for potential seasonal patterns in loan volume (KrediHacmi). Assuming there's a monthly seasonality with a cycle of roughly 20 weeks, I created two new features in the kredi data frame:

- seasonality_sin: This variable contains sine values based on the week number. The sine function introduces a wave-like pattern that repeats every 20 weeks, potentially capturing the rise and fall of loan volume throughout the year.
- seasonality_cos: This variable contains cosine values calculated similarly using the week number. The cosine function complements the sine function, and together they provide a more comprehensive representation of seasonality.

Fitting a Model with Seasonality:

I built a new linear regression model named l_fit_wls_fourier to see if including seasonality improves the model's ability to predict loan volume. This model incorporates the following predictors:

- MevduatFaizOrani: Deposit interest rate (from previous models)
- trnd: Time trend (from previous models)
- seasonality_sin: The newly created sine-based seasonality feature
- seasonality_cos: The newly created cosine-based seasonality feature

I used the weights I calculated earlier (to address potential heteroscedasticity) and the kredi data frame that now includes the seasonality features. I then examined the model fit using summary to see how the inclusion of seasonality affects the model's performance.

Evaluating the Model with Seasonality:

Ahmet Çeliker
2020402126

I checked the residuals of the model with seasonality (l_fit_wls_fourier) for normality using checkresiduals. Ideally, the residuals should be randomly scattered around zero if the model's assumptions hold.

Next, I created a visualization using ggplot. The plot shows the actual loan volume (KrediHacmi) on the y-axis and deposit interest rate (MevduatFaizOrani) on the x-axis. It includes points representing the data (geom_point) with some transparency (alpha adjusted).

More importantly, I overlaid a smooth line (geom_smooth) that represents the fitted model with WLS, trend, and the newly incorporated Fourier features (sine and cosine seasonality). I adjusted the color and line type for better visualization. The title clarifies that this is a fitted model with these additional features.

Checking Model Assumptions:

Now, I focused on verifying the assumptions of linear regression for the model with seasonality (l_fit_wls_fourier).

- I calculated the mean of the residuals. In an ideal scenario, the mean should be close to zero, indicating the model's predictions are unbiased.
- I checked for autocorrelation in the residuals using acf (autocorrelation function) and pacf (partial autocorrelation function). Ideally, there should be minimal autocorrelation to ensure the independence of errors in the model.
- I summarized the model again using summary for reference.
- Finally, I visually inspected the residuals again using checkresiduals.

Including Lags of Residuals (Optional):

This step depends on the results of the previous checks. If there's still significant autocorrelation in the residuals, I might need to take further action.

Here, I fit another model (l_fit_wls_fourier_ar) that includes all the previous predictors (deposit interest rate, trend, seasonality features) along with lags of the residuals from the previous model (l_fit_wls_fourier). The specific lags I used (here, lag 1 and potentially lag 2) might need adjustments based on the autocorrelation analysis.

The purpose of including lagged residuals is to capture any remaining serial dependence in the errors that the other features might not have addressed entirely.

Finally, I checked for autocorrelation in the residuals of this new model (l_fit_wls_fourier_ar) using acf and pacf to see if including lagged residuals helped.

By incorporating seasonality features and potentially addressing autocorrelation in the residuals, I aimed to improve the linear regression model for predicting loan volume. I'll need to analyze the results of these steps to assess the effectiveness of the introduced changes.

Ahmet Çeliker
2020402126

Research Model 2

In my analysis, I aim to understand the factors influencing car prices by examining variables such as credit interest rates, production quantity.

Specting dataset structure:

Using the str() function, I examined the structure of the dataset to understand its variables, data types, and dimensions. This helps in identifying potential issues with data types or missing values, ensuring data quality.

Creating a new dataframe:

I created a new dataframe arabafiyat2 by excluding the 'Tarih' column. This was done to prepare the data for correlation analysis and visualization. Excluding irrelevant columns streamlines the analysis and focuses on variables of interest.

Correlation analysis and visualization:

Calculating the correlation matrix using the cor() function helps in understanding the relationships between variables. Visualizing the correlation matrix with a heatmap using ggcorrplot provides a clear overview of variable associations, guiding further analysis.

Scatter plots and pair plots:

Scatter plots visualize the relationship between two continuous variables ('KrediFaizi' and 'ArabaFiyat'). They help in identifying patterns or trends in the data, such as linear or non-linear associations. Pair plots extend this analysis by plotting multiple variables in a matrix, allowing for quick examination of variable distributions and correlations.

Linear regression modeling:

Fitting a linear regression model (l_fit) allows for quantifying the relationship between predictors (such as 'KrediFaizi' and 'UretimAdet') and the response variable ('ArabaFiyat'). This helps in understanding how changes in predictors influence the outcome variable. The model summary provides insights into the significance of predictors and overall model fit.

Visualization of fitted model:

Plotting diagnostic plots (such as residuals versus fitted values) helps in assessing the assumptions of the linear regression model. These plots allow for identifying any patterns or trends in model residuals, ensuring the validity of the model.

Logarithmic transformation:

Logarithmic transformation of the response variable ('ArabaFiyat') and predictor ('KrediFaizi') is performed to address issues such as heteroscedasticity or non-linearity. This transformation can stabilize variance and linearize relationships, improving the performance of the regression model.

Time series visualization:

Plotting time series data for variables like 'ArabaFiyat', 'KrediFaizi', and 'UretimAdet' over time helps in identifying trends, seasonality, or other patterns in the data. This is crucial for understanding the temporal dynamics of the variables and their potential impact on the research question.

Initial Linear Regression Model:

- l_fit = lm(ArabaFiyat~.,data=arabafiyat2)**:** Here, I fit a linear regression model using the lm function from the base R package. The formula ArabaFiyat~. indicates that we're using all other variables in arabafiyat2 as independent variables to predict car sales prices (ArabaFiyat).
- summary(l_fit)**:** This line displays the summary of the fitted model, including coefficients, R-squared value (goodness-of-fit), and p-values. The coefficients tell you how much each independent variable is expected to influence car sales prices, while R-squared indicates the proportion of variance

Evaluating the Linear Regression Model:

- plot(l_fit)**:** This line creates a diagnostic plot of the fitted model. It typically shows the fitted regression line overlaid on the scatter plot of the actual data points. Examining this plot can reveal potential issues like non-linearity or outliers.
- ggplot(arabafiyat2, aes(x=KrediFaizi,y=ArabaFiyat)) + geom_point()+ geom_smooth(method='lm')**:** Here, I recreate the scatter plot between credit interest rates and car sales prices. Additionally, I add a smooth line using the geom_smooth function with the method='lm' argument. This overlaid line represents the fitted linear regression model, allowing you to visually assess how well it captures the trend in the data.
- ggplot(arabafiyat2, aes(x=KrediFaizi,y=ArabaFiyat)) + geom_point()+ geom_smooth(method='loess')**:** This line is similar to the previous one, but it uses geom_smooth(method='loess') instead. Loess is a non-parametric regression technique that can be more flexible in capturing non-linear relationships compared to the linear regression model. By comparing these two plots, you can see if a linear model is sufficient or if a more flexible approach might be necessary.

We repeat these steps for another variable, production quantity (UretimAdet), to explore its relationship with car sales prices.

- **ggplot(arabafiyat2, aes(x=UretimAdet,y=ArabaFiyat)) + geom_point()+ geom_smooth(method='lm')
- **ggplot(arabafiyat2, aes(x=UretimAdet,y=ArabaFiyat)) + geom_point()+ geom_smooth(method='loess')

Time Series Analysis of Car Sales Prices:

- ggplot(arabafiyat, aes(x = Tarih, y = ArabaFiyat)) + geom_line(color = "blue") + labs(title = "Araba Satış Fiyatı", x = "Tarih", y = "Satış Fiyatı")**:** This line creates a time series plot using ggplot2. It shows car sales prices (ArabaFiyat) on the y-axis and the time variable (Tarih) on the x-axis. This plot allows you to visualize trends and seasonality patterns in car sales prices over time.

Logarithmic Transformation (Attempted but Not Used):

- arabafiyat$ArabaFiyat <- log(arabafiyat$ArabaFiyat + 1)**:** Here, I attempt a logarithmic transformation on the car sales prices. The log(x + 1) ensures we don't take the logarithm of zero. This transformation can sometimes be helpful for stabilizing the variance or improving normality of the data.
- summary(arabafiyat)**:** I use summary to check the summary statistics after the transformation.
- l_fit_log <- lm(ArabaFiyat~., data=arabafiyat)**:** I fit a new linear regression model using the log-transformed car sales prices.
- summary(l_fit_log)**:** The summary of the new model is displayed.
- ggplot(arabafiyat, aes(x = Tarih, y = ArabaFiyat)) + geom_line(color = "blue") + labs(title = "Araba Satış Fiyatı Log Transformed", x = "Tarih", y = "Satış Fiyatı Log")**:** This line creates a time series plot for the log-transformed car sales prices.
- ggplot(arabafiyat, aes(x = Tarih, y = KrediFaizi)) + geom_line(color = "blue") + labs(title = "Araba Kredi Faizi", x = "Tarih", y = "KrediFaizi")**:** This plot shows the time series of credit interest rates.
- logtransformed = copy(arabafiyat)**:** I create a copy of the data frame with the log-transformed car sales prices.
- logtransformed$KrediFaizi <- log(logtransformed$KrediFaizi + 1)**:** Here, I apply a log transformation to the credit interest rates as well.
- summary(logtransformed)**:** I check the summary statistics after the transformation on credit interest rates.
- head(logtransformed)**:** This line displays the first few rows of the transformed data frame to visually inspect the changes.

Checking for Stationarity (Initial Attempt):

- **"After applying the logarithmic transformation, there wasn't a significant change, so there's no need to take the logarithm of the credit interest rate data."** Based on the results (not explicitly shown), I concluded that the log transformation didn't significantly improve the data, so I decided to proceed with the original credit interest rate values.

Time Series Plots of Other Variables:

- ggplot(arabafiyat, aes(x = Tarih, y = UretimAdet)) + geom_line(color = "blue") + labs(title = "Araba Uretim Adedi", x = "Tarih", y = "Uretim")**:** This line creates a time series plot for the production quantity (UretimAdet), allowing you to visualize trends and seasonality patterns in car production over time.

Logarithmic Transformation on Production Quantity (Attempted but Not Used):

- logtransformed$UretimAdet <- log(logtransformed$UretimAdet + 1)**:** Similar to car sales prices, I attempt a log transformation on the production quantity.
- summary(logtransformed)**:** I check the summary statistics to see the impact of the transformation.
- head(logtransformed)**:** This line displays the first few rows of the transformed data frame for inspection.

Including Time Trend:

- library(dplyr)**:** I load the dplyr library, which provides functions for data manipulation.
- logtransformed <- logtransformed %>% mutate(trnd = row_number())**:** Here, I add a new variable named trnd to the logtransformed data frame. This variable simply assigns a sequential number to each row, essentially creating a time trend variable. This can be useful for capturing the overall increasing or decreasing trend in car sales prices over time, along with the effects of other variables.

- om_line(color = "blue") + labs(x = "Tarih", y = "Satış Fiyatı")**:** This line creates a time series plot of the actual car sales prices (ArabaFiyat) along with the predicted values from the model (l_fit_log_diff - the model with the trend, assuming it's the one currently assigned). This allows you to visually assess how well the model captures the trend and fluctuations in car sales prices over time.

Building the Initial Model:

I included a time trend variable (trnd) to capture the overall increasing or decreasing trend in car sales prices. The initial time series regression model was built using lm to predict log-transformed car sales prices based on credit interest rates, production quantity, and the trend.

Checking Model Assumptions:

I assessed the model's residuals for normality, homoscedasticity, and stationarity using various functions from lmtest and tseries libraries. Additionally, I visually inspected the residuals over time and checked for autocorrelation using acf and pacf.

Addressing Potential Issues:

I identified and removed potential outliers based on my analysis of the data. This could improve the model's fit by reducing the influence of extreme values.

Refining the Model:

I refit the model with the improved data (potentially excluding outliers) and examined the relationships between variables using scatter plots.

Weighted Least Squares Regression:

Ahmet Çeliker
2020402126

To address potential heteroscedasticity, I implemented Weighted Least Squares (WLS) regression. This assigns higher weights to data points with smaller residuals, potentially improving the model's fit.

Evaluating the WLS Model:

I assessed the WLS model's residuals for randomness and lack of autocorrelation, compared to the unweighted model. Finally, I visualized the actual car sales prices alongside the predicted values from the WLS model to see how well it captured the trends and fluctuations in the data.

This analysis provided valuable insights into the factors influencing car sales prices and allowed me to build a time series regression model that considers these factors and potential issues.

Conclusion

This analysis investigated the factors influencing car sales prices in a provided dataset. Exploratory data analysis revealed relationships between car sales prices, credit interest rates, and production quantity. Time series plots visualized trends and seasonality patterns in car sales prices.

A time series regression model was built to predict car sales prices based on credit interest rates, production quantity, and a time trend variable. The model's residuals were analyzed to ensure they met the assumptions of normality, homoscedasticity, and stationality. Potential outliers were identified and removed to improve the model's fit.

Weighted Least Squares regression was implemented to address potential heteroscedasticity in the residuals. The final model's residuals were assessed, and the predicted car sales prices were visualized alongside the actual data.

Research Model 3

Research Question

In this analysis, I aim to investigate the relationship between credit card spending (Kredi Kartı Harcama) and money supply (Para Arzı) in Turkey, while also considering the influence of electricity consumption (Elektrik Tüketim). My research question is: "Are changes in credit card spending related to fluctuations in money supply and electricity consumption in Turkey?"

Converting date format:

The 'Tarih' column typically contains date information in a specific format. However, R may not recognize it as a date object initially. Therefore, we use the **convert_to_ymd()** function to convert the date string to the appropriate date format.

Converting the date format ensures uniformity in date representation, facilitating time series analysis and visualization.

Visualizing time series data:

Visualization is a crucial step in exploratory data analysis. By plotting each time series variable over time, we can visually inspect the trends, patterns, and fluctuations in the data.

Line plots are chosen for visualization as they effectively represent the changes in each variable over time. Each line represents a variable, making it easy to compare their behaviors.

Insights gained from visualizations can include identifying trends (upward, downward, or stationary), seasonality, and any unusual patterns or outliers in the data.

Correlation analysis:

Correlation analysis helps in understanding the relationships between variables by quantifying the strength and direction of linear associations.

Computing the correlation matrix provides an overview of pairwise correlations between credit card spending, money supply, and electricity consumption.

Visualizing the correlation matrix with a heatmap makes it easier to identify strong correlations (positive or negative) between variables. A heatmap color-codes the correlation values, highlighting significant associations.

Insights from correlation analysis can guide further investigation into potential causal relationships or dependencies between variables.

Time Series Conversion:

The Tarih (Date) column initially appears as text. To perform time-based analysis, I converted it to a proper date object using a function convert_to_ymd. This function uses as.Date to ensure dates are recognized correctly for further analysis.

Line Plots:

I created individual line plots for each variable (KrediKartıHarcama (Credit Card Spending), ParaArzı (Money Supply), and ElektrikTüketim (Electricity Consumption)) using ggplot. Visualizing these trends over time (x-axis being Date) allows for initial inspection of patterns and potential relationships.

Ahmet Çeliker
2020402126

Correlation Analysis:

To explore how the numerical variables are related, I calculated the correlation coefficients using the cor function. This resulted in a correlation matrix, correl_info.

I then utilized ggcorrplot to create a correlation heatmap, which visually represents the strength and direction of the relationships between each pair of variables. The heatmap helps identify potential explanatory variables for credit card spending.

Additionally, I used ggpairs to generate a scatterplot matrix. This provides a more detailed view of the pairwise relationships between all numerical variables, allowing for a closer examination of potential linear or non-linear associations.

Time Series Regression Analysis

To explore how other factors might influence credit card spending, I built linear regression models. My initial model included credit card spending as the dependent variable and all other numerical variables in the data as independent variables. The model summary (summary(l_fit)) provided information about the coefficients, p-values, and R-squared value. These helped me assess the direction, strength, and significance of the relationships between each variable and credit card spending.

Considering that linear regression often performs better with linear relationships, I applied logarithmic transformations to some variables (Money Supply, Credit Card Spending). I then built a new regression model (l_fit_log) using these transformations alongside the other explanatory variables. By analyzing the model summary, visualizations, and residuals, I assessed the impact of the transformations on the model's fit.

Incorporating Trends and Categorical Variables

To capture potential trends in spending behavior, I incorporated rolling averages for credit card spending over the past 3 and 6 months (Trend_3_Aylik, Trend_6_Aylik) as new features. Additionally, I included month (ay) and year (yil) as categorical variables to account for possible seasonal effects or economic changes over time. A new regression model (l_fit_trend_cathegorical) was built using these features alongside the existing ones. The model's fit was evaluated using the summary, visualizations, and residual checks to determine if these additions improved the model's explanatory power.

Lagged Variables and Model Refinement

The possibility of delayed effects from some variables (e.g., Money Supply) on credit card spending was addressed by introducing lagged credit card spending variables. These variables represented credit card spending values from previous months (e.g., 4 months ago). I built new regression models (l_fit_gecikmeli) with these lagged variables and the logarithmic transformations. By comparing the model summaries and residuals of these models to previous ones, I aimed to identify if including lagged variables led to a better explanation of credit card spending patterns.

Ahmet Çeliker
2020402126

Conclusion

This analysis employed time series data and regression modeling to explore factors influencing credit card spending in Turkey. The initial models provided a baseline understanding of the relationships between variables. Subsequent models with logarithmic transformations, trend variables, categorical variables, and lagged variables aimed to improve the model's ability to explain credit card spending variations. By comparing and evaluating these models, I gained valuable insights into the factors that might drive credit card spending behavior in Turkey.