MŰEGYETEM 1782

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Computer Science and Information Theory

Assignment (#15)

Written by: Cuadros Rivas, Alejandra Paola **Neptun Code:** KK5459

Subject: Mathematical Statistics

E-mail: acuadrosrivas@edu.bme.hu

Responsible Lecturer : Dr. Pintér, Márta

Academic year: 2023/1 Fall semester

Problem 1

A study by a shoe brand collected the effect of new shoes on people's walking speed and self-confidence. The data were recorded in file 15assign1.csv

Questions

We wish to give a linear regression model, in which the response variable is self-confidence, while the explanatory variables are the number of days the shoes have been worn and the walking speed.

1. **Estimate and interpret the parameters of the linear regression, test whether the explanatory variables are significant (at significance level 5%).**

   Statistical Method:

   Linear regression using "lm()".

```
Call:
lm(formula = Self_confidence ~ Shoe_wearing_days + Walking_speed_km
_h, data = X15assign1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.053006 -0.035537  0.001931  0.005105  0.052551 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6.7639878  0.0362682  186.50   <2e-16 ***
Shoe_wearing_days   0.2499868  0.0003574  699.45   <2e-16 ***
Walking_speed_km_h -0.0048838  0.0113637   -0.43    0.669    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03638 on 47 degrees of freedom
Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999 
F-statistic: 2.458e+05 on 2 and 47 DF,  p-value: < 2.2e-16
```

   Linear Regression Model:

   Self-confidence = 6.7639878 + 0.2499868 × Shoe wearing days − 0.0048838 × Walking speed (km/h)

   Interpretation:

   a. The **Residuals**, that are the differences between observed and predicted values, are in a range about -0.053 to 0.053, it has a median very close to 0 (0.0019). The model fits the data well, with not systematic bias in predictions.

   b. About the **Coefficients:**

      a. The intercept (6.764) is the estimated value when "Shoe wearing days" and "Walking speed (km/h)" are 0. The high result suggests a baseline level of self confidence in the absence of these factors.

   b. Shoe wearing days (0.25) indicates that for each additional day of shoe wearing, the score increases in 0.25 points.
   c. Walking speed (km/h) (-0.005) suggests that an increase in walking speed is associated with a slight decrease in "Self-confidence", although that relationship is not statistically significant due to the p-value is 0.669.

c. The **Residual Standard Error** (0.0364) is the average amount the response will deviate from the true regression line, as its small, this indicates good model fit.
d. **R-squared** (0.9999) indicates the proportion of variance in Self Confidence, suggesting an excellent fit cause it's really high, also it may indicate overfitting.
e. The **F-statistic (2.458e+05)** and **p-value(< 2.2e-16)**, determine the importance of the model, the high F-statistic and slow p-value indicate that is statistically significant.

Significance of explanatory variables at significance level 5%

The variable "Shoe wearing days" is a significant predictor of "Self-confidence" cause has a p-value much lower than 0.05, while "Walking speed (km/h)" is not cause has a p-value higher than 0.05.

2. **Define and interpret the multivariate coefficient of determination.**

Statistical Method:

R^2 from the model summary.

```
> summary(model)$r.squared
[1] 0.9999044
```

Interpretation:

The model explains 100% of the variance in self-confidence, which is unusually high and might indicate overfitting.

3. **Test the reliability of the regression model at significance level 5% .**

Statistical Method:

F-test from the model summary.

```
F-statistic: 2.458e+05 on 2 and 47 DF,  p-value: < 2.2e-16
```

Interpretation:

The p-value is extremely low compared to 5%, this suggests that the model is statically significant at the 5% significance level.

4. **Give an interval estimate for the parameters with 95% confidence level.**

Statistical Method:

95% Confidence Intervals for model parameters

```
> confint(model, level = 0.95)
                        2.5 %     97.5 %
(Intercept)        6.69102558 6.83694996
Shoe_wearing_days  0.24926780 0.25070582
Walking_speed_km_h -0.02774467 0.01797704
```

Interpretation

a. Intercept [6.69102558, 6.83694996] suggests that we are 95% confident that the true value of the intercept will lie between the Confidence interval when both parameters are 0.

b. Shoe wearing days [0.24926780, 0.25070582] suggesting that we are 95% confident that the true effect of an additional shoe wearing on “self conficence” is within the range. There is a high level of precision in the estimate of the coefficient due to the narrow range.

c. Walking Speed (km/h) [-0.02774467, 0.01797704] suggests that the effect of walking speed on “self-confidence” is uncertain and may not be statistically significant, due to the zero in the interval.

5. Give an estimation for the if the walking speed is 5.3 km/h and the shoes have been worn for 9 days. Also, give a 95% confidence interval estimate for that.

Statistical Method:

Prediction with confidence intervals

```
> new_data <- data.frame(Shoe_wearing_days = 9, Walking_speed_km_h
= 5.3)
> prediction <- predict(model, newdata = new_data, interval = "pred
iction", level = 0.95)
>
> print(prediction)
       fit      lwr      upr
1 8.987985 8.896983 9.078987
```

Interpretation:

When the walking speed is 5.3 km/h and the shoes have been worn for 9 days, with a 95% of confidence interval:

Under those conditions, the **Fit** is 8.9880 and the **Confidence interval** is [8.8970, 9.0790]. This reflects the high level of certainty about the prediction.

Problem 2

The effect of another shoe brand on the walking speed of people in different age groups was investigated. File 15assign2.csv contains the collected data.

Question

1. **Test whether there is a significant difference between walking speeds as a result of the shoes in the different age groups at significance level ε = 0.05.**

Statistical Method:

ANOVA(Analysis of Variance), we use this method cause by the dataset we can assume that:

- The samples from different groups are **independent**.
- The data in each group comes from a **normally** distributed population.
- The variances of the population are equal, so there's **homogeneity.**

H0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H1: At least one μ_i is different

```
> X15assign2$id <- seq_len(nrow(X15assign2))
> X15assign2_Reshaped <- melt(X15assign2, id.vars = "id", variable.n
ame = "age_group", value.name = "walking_speed")
> X15assign2_Reshaped$id <- seq_len(nrow(X15assign2_Reshaped))
> result <- aov(walking_speed ~ age_group, data=X15assign2_Reshaped)
> summary(result)
             Df Sum Sq Mean Sq F value Pr(>F)    
age_group     3 10.412   3.471   103.2 <2e-16 ***
Residuals   156  5.248   0.034                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

a. Degrees of Freedom (Df): The age_group variable represents different age groups has 3 degrees of freedom. There are 156 degrees of freedom for the residuals.
b. Sum of Squares (Sum Sq): For the age_group is 10.412 and for residuals is 5.248
c. Mean Square (Mean Sq): For age_group is 3.471 and for residuals is 0.034.
d. F value: 103.20 indicates a more significant effect of the factor being tested.
e. Pr(>F): is less than 2e-16, notation for a value extremely close to zero, significantly lower than the typical significance level of 0.05
f. Significance: *** next to the p-value indicates a very high level of statistical significance.

- The very low p-value (<2e-16) suggests that there are statistically significant differences in walking speeds across different age groups and evidence against the null hypothesis.

- The large F value (103.2) reinforces this finding, indicating a strong effect of age group on walking speed.
- Given this result, it would be appropriate to conduct other tests:

Checking Normality with a Shapiro-Wilk test

```
# Q-Q plot of residuals
> qqnorm(residuals(result))
> qqline(residuals(result))
>
> # Shapiro-Wilk normality test
> shapiro.test(residuals(result))

        Shapiro-Wilk normality test

data:  residuals(result)
W = 0.95721, p-value = 7.919e-05
```

The test statistic is W = 0.95721. The Shapiro-Wilk test statistic ranges from 0 to 1, where a value closer to 1 indicates that the data is more normally distributed.

The p-value is 7.919e-05, which is much smaller than the conventional alpha level of 0.05. A small p-value (typically ≤ 0.05) indicates that **the null hypothesis of normality can be rejected.**

The Shapiro-Wilk test results suggest that **the residuals from the ANOVA model do not follow a normal distribution.**

Checking Homogeneity of Variances

```
> leveneTest(walking_speed ~ age_group, data = X15assign2_Reshaped)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   3   4.875 0.002873 **
      156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F value is 4.875. This is the test statistic for Levene's test and indicates the ratio of the variance between the groups to the variance within the groups. The p-value is 0.002873, which is less than the conventional significance level of 0.05.

The p-value indicates that there are statistically significant differences in variances across the groups. Since the p-value is below 0.05, this suggests that **the assumption of homogeneity of variances is violated**.

When this assumption is violated, the results of the ANOVA may not be reliable, let's try with Welch's ANOVA.

Welch's ANOVA: This is a variation of ANOVA that is more robust to violations of the homogeneity of variances.

```
> oneway.test(walking_speed ~ age_group, data = X15assign2_Reshaped)

	One-way analysis of means (not assuming equal variances)

data:  walking_speed and age_group
F = 161.37, num df = 3.000, denom df = 85.225, p-value < 2.2e-16
```

The F statistic is 161.37, which is a measure of the ratio of the variance between the groups to the variance within the groups. The p-value is less than 2.2e-16, which is extremely small.

Interpretation:

Given that the p-value is much less than the significance level ($\varepsilon = 0.05$), you would reject the null hypothesis (H0) that states there is no significant difference in walking speeds among the different age groups due to the shoes.

The results suggest that there is a statistically significant difference in walking speeds across the different age groups when wearing the shoes in question.

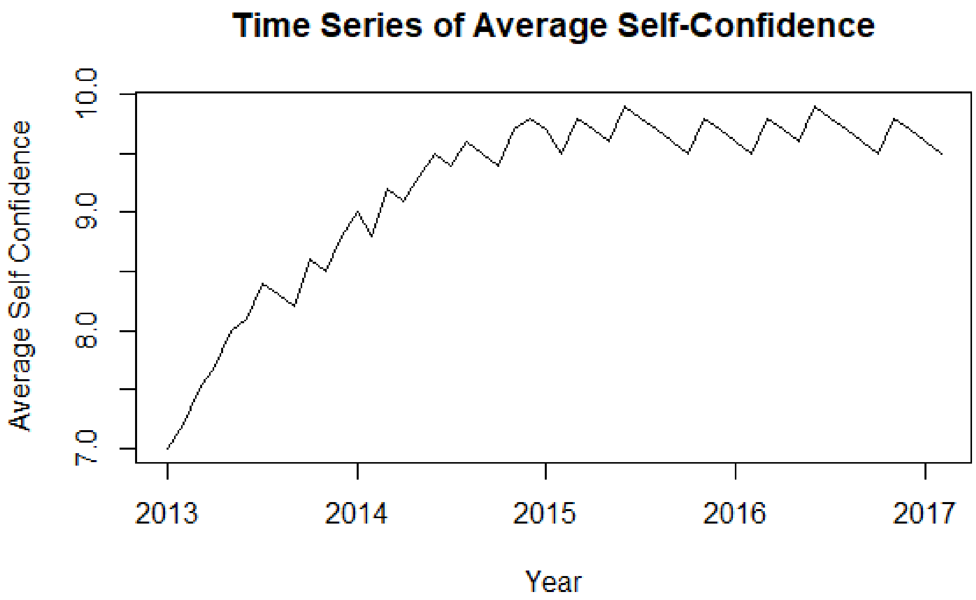
Problem 3

File 15assign3.csv shows a survey about the effect of new shoes on people's self-confidence in a few years/months.

Questions

1. **Make a time series diagram based on the data, then calculate the empirical autocorrelation and partial autocorrelation functions.**
   a. **Statistical Method:**
      - **Time Series Plot** to visualize the data over time.

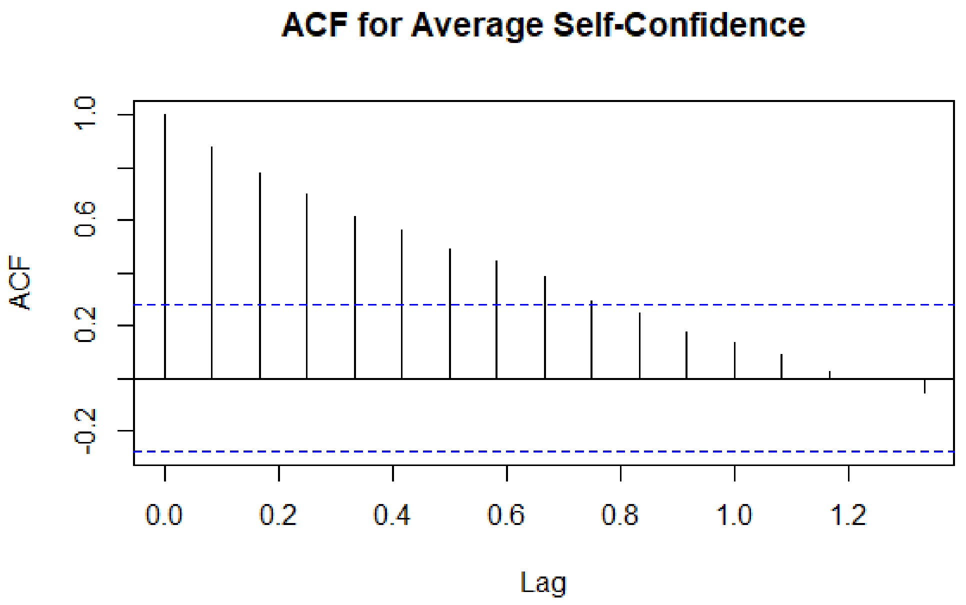
```
> X15assign3$Date <- ymd(paste(X15assign3$`Year_Month`, "01", sep =
 "-"))
> # Time Series Plot
> X15assign3_ts <- ts(X15assign3$Average_Self_Confidence, start=c(2
013,1), frequency=12)
> plot(X15assign3_ts, main='Time Series of Average Self-Confidence'
, xlab='Year', ylab='Average Self Confidence')
```



Interpretation: The time series show how self-confidence changes over time, showing trends and fluctuations of the data. Theres an upward trend, this one suggests an increase in average self-confidence over the years

   b. **Statistical Method:**
      - **Autocorrelation Function (ACF)** to measure how the observations are correlated with each other at different lags.

```
> acf(X15assign3_ts, main='ACF for Average Self-Confidence')
```

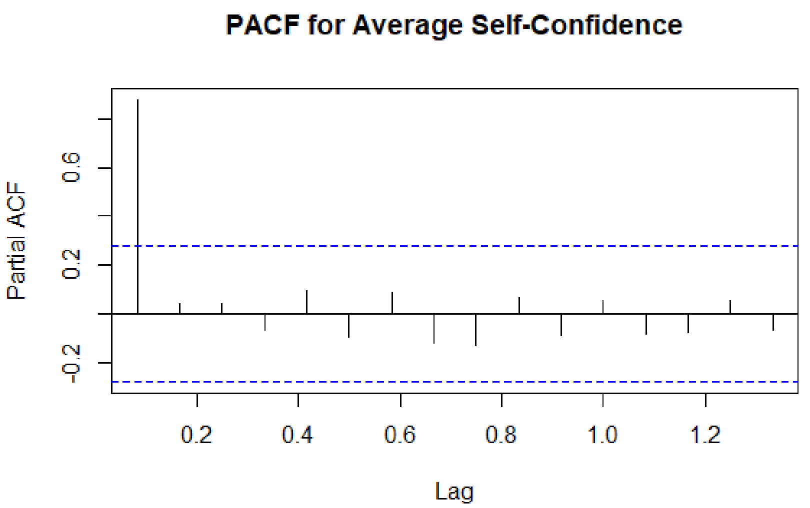


Interpretation: ACF shows the correlation of the series with itself at different lags, the peaks in the plot indicate a significant correlation at specific lags, which suggests is seasonality. There are a few lags that have significant autocorrelation since the bars extend beyond the blue confidence interval bounds. This indicates that past values have some correlation with future values. The slow decay of the autocorrelation suggests that the series may be non-stationary.

c. **Statistical Method:**

- **Partial Autocorrelation Function (PACF)** To measure the correlation between observations at different lags, controlling for the values of the observations at all shorter lags.

```
> pacf(X15assign3_ts, main='PACF for Average Self-Confidence')
```



Interpretation: PACF plot displays the correlation of the series with its lags. This plot shows a significant spike at the first lag, which quickly falls off in subsequent lags. This indicates that an AR(1) model might be a good starting point for modelling this time series, as there is a significant correlation with the first lag that isn't explained by the others.

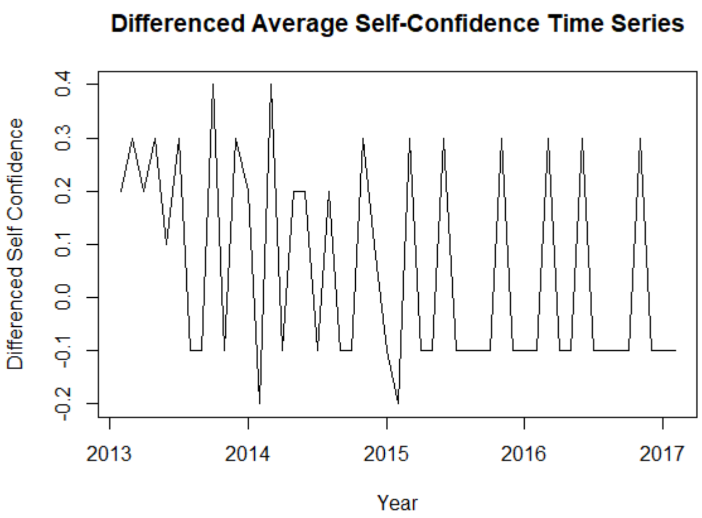
2. **Experiment with transforming the data and filtering out the trend and seasonal components to fit different time series models. Perform fit (relevance) testing.**

Statistical Method:

a. **Data Transformation** to experiment with different transformations like differencing, logarithmic and square root to stabilize the variance and make the time series more stationary.

Differenced Average Self-Confidence Time Series

```
> x15assign3_diff <- diff(x15assign3_ts)
> plot(x15assign3_diff, main='Differenced Average Self-Confidence T
ime Series', xlab='Year', ylab='Differenced Self Confidence')
```

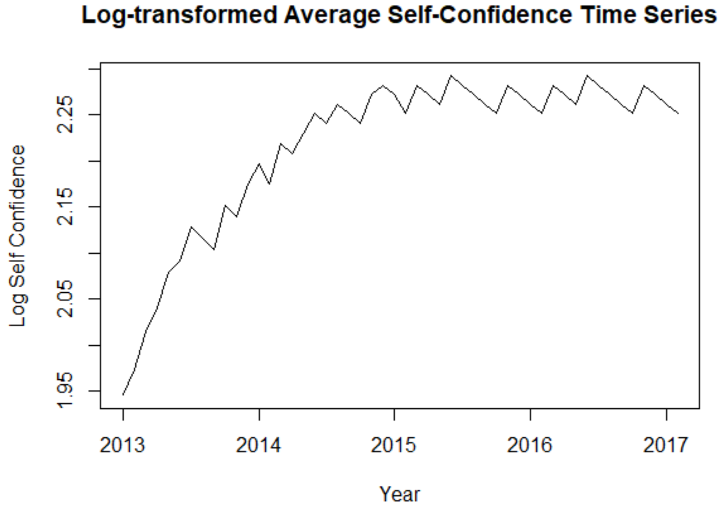


Interpretation:
The differencing appears to have achieved stationarity, as the mean seems constant over time, which is a key requirement for many time series models. The plot shows fluctuations around a mean of zero without any apparent trend or seasonal patterns, suggesting that differencing has effectively removed these components.

The lack of trend and seasonality implies that the differenced series could be well-represented by an ARIMA model. There's no clear evidence of seasonality in the differenced series.

Log-Transformed Average Self-Confidence Time Series

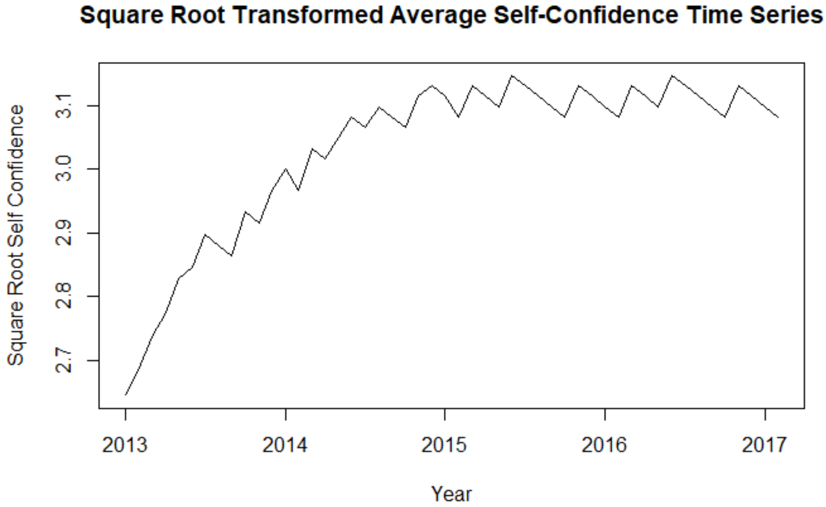
```
> X15assign3_log <- log(X15assign3_ts)
> plot(X15assign3_log, main='Log-transformed Average Self-Confidenc
e Time Series', xlab='Year', ylab='Log Self Confidence')
```



Interpretation:
The log transformation aims to stabilize the variance, which is beneficial when the original series exhibits exponential growth or variance increasing with the level. The data shows a clear upward trend until early 2016, after which it remains constant. This indicates a period of growth in self-confidence followed by a levelling off. The presence of a trend suggests that further differencing may be required to achieve stationarity, there's no obvious seasonality.

Square Root Transformed Average Self-Confidence Time Series

```
> X15assign3_sqrt <- sqrt(X15assign3_ts)
> plot(X15assign3_sqrt, main='Square Root Transformed Average Self-
Confidence Time Series', xlab='Year', ylab='Square Root Self Confid
ence')
```

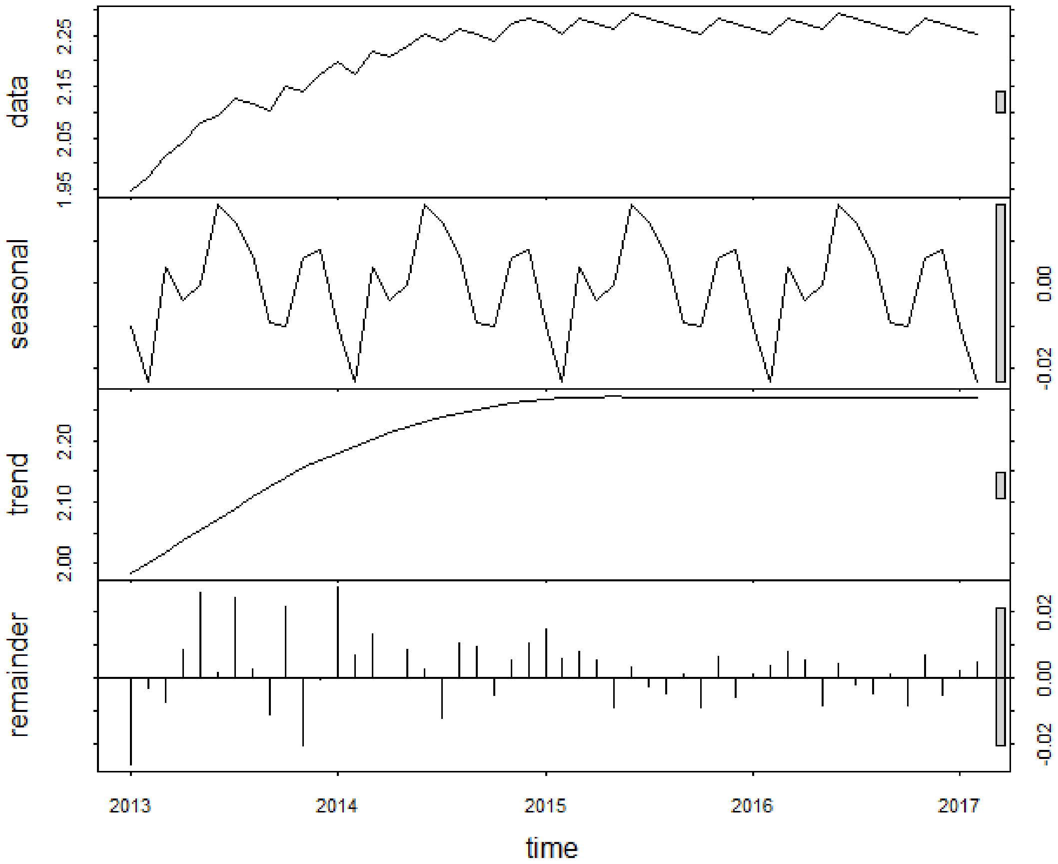


Interpretation:
An increasing trend is observed until late 2015, after which the self-confidence levels are maintained. The square root transformation has likely made the variance more uniform across the time series. The data may still need differencing to achieve stationarity, as indicated by the persisting trend. No clear seasonal pattern is visible.

b. Decomposing Time Series to observe the trend, seasonality, and random components.

Decomposing the Log-Transformed Series

```
> X15assign3_log_decomposed <- stl(X15assign3_log, s.window="period
ic")
> plot(X15assign3_log_decomposed)
```



Interpretation:

There is a distinct and consistent seasonal pattern, suggesting that self-confidence fluctuates at regular intervals throughout the year.

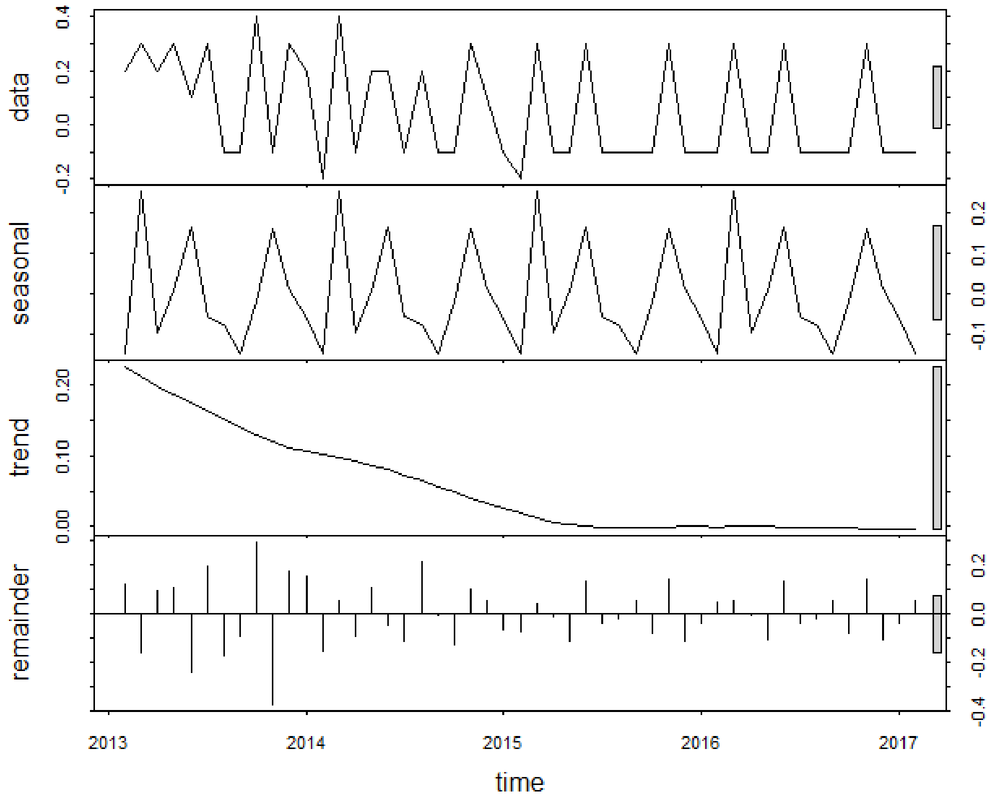
An initial increasing trend in self-confidence levels off in later years, indicating a period of growth followed by stabilization.

The residuals show no apparent pattern post-decomposition, indicating that the trend and seasonal components have been effectively captured.

This analysis suggests the potential utility of a seasonal ARIMA model to forecast future values, accounting for both the identified trend and seasonality.

Decomposing the Differenced Average Series

```
> X15assign3_diff_decomposed <- stl(X15assign3_diff, s.window="peri
odic")
> plot(X15assign3_diff_decomposed)
```



Interpretation:

The seasonal component displays a consistent pattern, suggesting that even after differencing, there is a recurring seasonal effect in the self-confidence data.

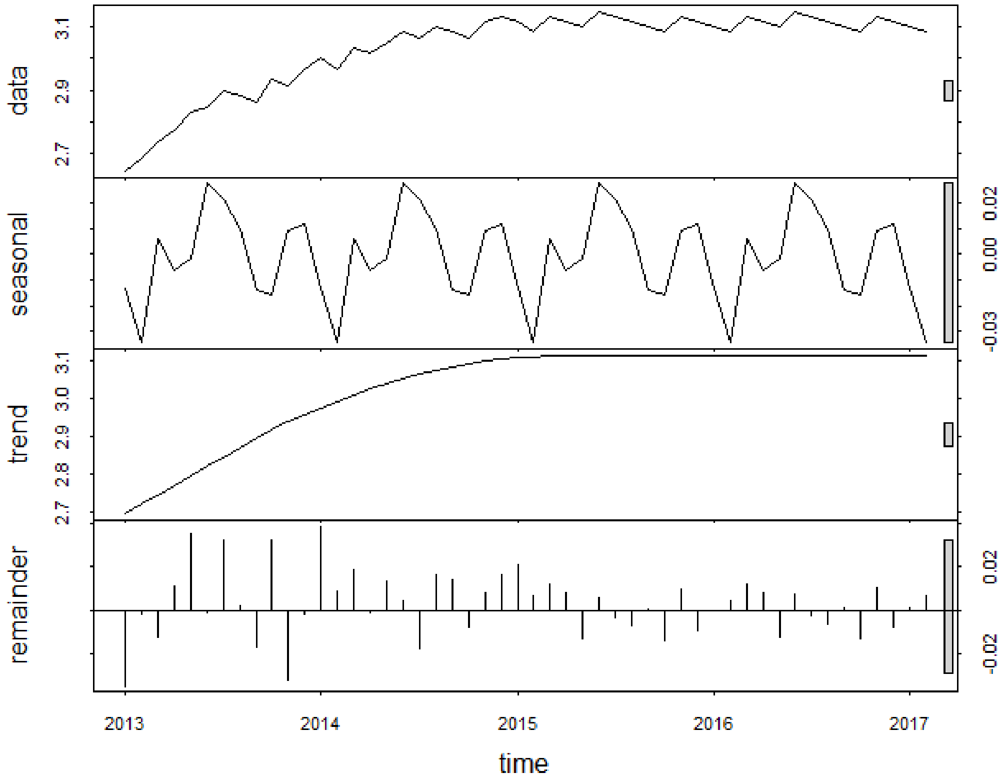
The trend component is relatively flat, indicating that the differencing has largely removed any trend from the data, as intended with differencing.

The residuals, which represent the remainder after seasonal and trend components are accounted for, show some variability but no clear patterns, indicating that the differenced series has randomness expected in stationary data.

This analysis would suggest that any models applied to the differenced data should account for the observed seasonality. Despite the differencing, the persistence of a seasonal pattern indicates that a seasonal model may be needed, such as SARIMA.

Decomposing the Square Root Transformed Average Series

```
> X15assign3_sqrt_decomposed <- stl(X15assign3_sqrt, s.window="peri
odic")
> plot(X15assign3_sqrt_decomposed)
```



Interpretation:

Shows the original square root transformed data with an evident upward trend and some oscillation that might indicate seasonality.

There's a clear, regular seasonal pattern, indicating that self-confidence has predictable fluctuations throughout each year.

A noticeable upward trend is present, which means that self-confidence increases over time before the series begins to plateau.

The residuals are mostly centred around zero with no apparent patterns, suggesting that the seasonal and trend components have been effectively captured by the decomposition.

These findings suggest that the square root transformation, followed by STL decomposition, has clarified the trend and seasonal structures in the data. For predictive modelling, a seasonal model that accounts for both the increasing trend and the clear seasonality, such as a SARIMA model, could be appropriate.

c. Model Fitting, fit different time series models to the transformed data.

ARIMA Model on Log-Transformed Data

```
> fit_log_arima <- auto.arima(X15assign3_log)
> summary(fit_log_arima)
Series: X15assign3_log
ARIMA(2,2,2)(1,0,0)[12]

Coefficients:
          ar1      ar2     ma1      ma2    sar1
      -1.2425  -0.7400  0.0464  -0.5889  0.7319
s.e.   0.1049   0.1063  0.1252   0.1140  0.0994

sigma^2 = 0.0001812:  log likelihood = 135.32
AIC=-258.63   AICc=-256.59   BIC=-247.41

Training set error measures:
                        ME       RMSE         MAE        MPE      MAPE
Training set -0.0006032116 0.01248449 0.009300015 -0.0329589 0.4190114
                  MASE        ACF1
Training set 0.1521606 -0.05837251
```

Interpretation: The ARIMA(2,2,2)(1,0,0)[12] model fitted on the log-transformed data effectively captures the underlying patterns with both seasonal and non-seasonal components. The model's low AIC and BIC values indicate a good fit, and the error metrics suggest accurate predictions with minimal bias. The near-zero ACF value for the residuals indicates that the model residuals are random, which is an indicator of a well-fitting model.

ARIMA Model on Differenced Average Data

```
> fit_diff_arima <- auto.arima(X15assign3_diff)
> summary(fit_diff_arima)
Series: X15assign3_diff
ARIMA(2,1,2)(1,0,0)[12]

Coefficients:
          ar1      ar2      ma1      ma2    sar1
      -1.2277  -0.7359  -0.0281  -0.6308  0.7405
s.e.   0.1058   0.1055   0.1130   0.1035  0.0949

sigma^2 = 0.01395:  log likelihood = 30.78
AIC=-49.57   AICc=-47.52   BIC=-38.34

Training set error measures:
                       ME      RMSE        MAE      MPE     MAPE      MASE
Training set -0.006803001 0.1106502 0.08304187 39.50717 59.53331 0.7315593
                    ACF1
Training set -0.09045192
Training set 0.1521606 -0.05837251
```

Interpretation: The ARIMA(2,1,2)(1,0,0)[12] model applied to the differenced average data is characterized by:

- A mix of autoregressive (AR) and moving average (MA) components, indicating a complex interaction of factors influencing self-confidence levels.
- The necessity for one level of differencing to achieve stationarity, which is less than what was needed for the log-transformed series.
- A seasonal autoregressive component that accounts for yearly patterns.
- Decent model fit, evidenced by relatively low AIC and BIC scores.

- The residuals' ACF value suggests a small amount of remaining autocorrelation after fitting the model.

The model appears to be a reasonable fit for the differenced data, capturing the essential dynamics and showing a capacity for making accurate predictions within a certain confidence interval.

ARIMA Model on Squared Root Transformed Data

```
> fit_sqrt_arima <- auto.arima(X15assign3_sqrt)
> summary(fit_sqrt_arima)
Series: X15assign3_sqrt
ARIMA(2,2,2)(1,0,0)[12]

Coefficients:
          ar1      ar2     ma1     ma2    sar1
      -1.2354  -0.7383  0.0046  -0.614  0.7373
s.e.   0.1052   0.1058  0.1174   0.107  0.0969

sigma^2 = 0.0003949:  log likelihood = 116.46
AIC=-220.92   AICc=-218.87   BIC=-209.69

Training set error measures:
                       ME       RMSE        MAE         MPE      MAPE
Training set -0.001023591 0.01842913 0.01375076 -0.04023967 0.4526332
                  MASE        ACF1
Training set 0.1523053 -0.07488807
```

Interpretation:

- The model structure is similar to that of the log and differenced data, suggesting consistent behaviour across transformations.
- Low AIC and BIC values indicate a good model fit for the square root transformed series.
- Residuals are well-behaved with low autocorrelation, suggesting the model captures the data's patterns effectively.

The model demonstrates adequate fitting, with low error metrics and good control of autocorrelation in the residuals, indicating potential for accurate forecasting.

d. **Model Selection and Fit Testing,** comparing the models using criteria like AIC or BIC and perform diagnostic checks on the chosen model.

Model Comparison

```
> AIC(fit_log_arima, fit_diff_arima, fit_sqrt_arima)
               df        AIC
fit_log_arima   6 -258.63404
fit_diff_arima  6  -49.56577
fit_sqrt_arima  6 -220.91843
```

Interpretation:

The AIC (Akaike Information Criterion) is used to compare models; the lower the AIC, the better the model balances fit and complexity.

fit_log_arima has the lowest AIC of -258.63, which suggests it provides the best fit among the three models with the least complexity, fit_diff_arima -49.57, less favorable compared to the log-transformed model, indicating a poorer fit or unnecessary complexity and fit_sqrt_arima -220.92, better than the differenced model but not as good as the log-transformed model.

Based on the AIC values, the fit_log_arima model is the preferred model for forecasting future values of the average self-confidence time series.

```
> BIC(fit_log_arima, fit_diff_arima, fit_sqrt_arima)
               df        BIC
fit_log_arima   6 -247.40683
fit_diff_arima  6  -38.33856
fit_sqrt_arima  6 -209.69123
```

Interpretation:

Based on the BIC (Bayesian Information Criterion),a lower BIC suggests that a model is better, as it implies either a higher likelihood or a simpler model with fewer parameters.

The log-transformed ARIMA model still performs the best with the lowest BIC of -247.41, reinforcing its status as the most suitable model among the three, the differenced data ARIMA model has the highest BIC of -38.34, suggesting it is less preferable due to either a poor fit or excessive complexity relative to its fit and the square root transformed ARIMA model has a BIC of -209.69, which is more favourable than the differenced model but not as optimal as the log-transformed model.

Diagnostic Checks

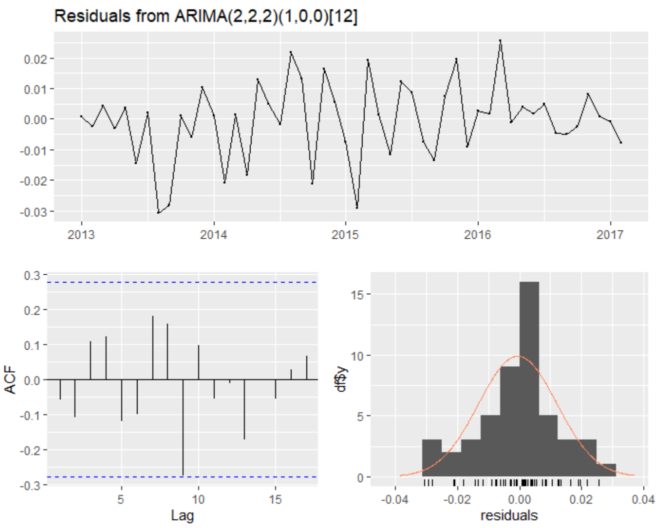
Diagnostic checks for the log-transformed ARIMA model

```
> checkresiduals(fit_log_arima)

        Ljung-Box test

data:  Residuals from ARIMA(2,2,2)(1,0,0)[12]
Q* = 12.595, df = 5, p-value = 0.02749

Model df: 5.   Total lags used: 10
```



Interpretation: This model is adequately fitted, with residuals that do not show significant patterns or bias and are close to normally distributed. However, the Ljung-Box test suggests that there might be some remaining autocorrelation not captured by the model.

Diagnostic checks for the Differenced Average ARIMA model

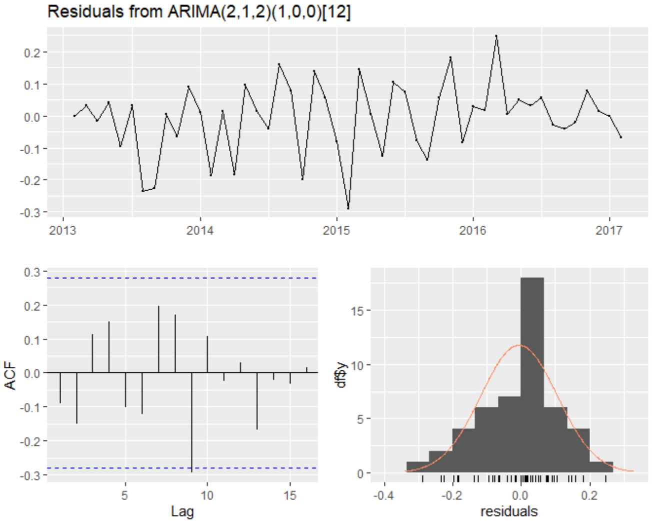
```
> checkresiduals(fit_diff_arima)

        Ljung-Box test

data:  Residuals from ARIMA(2,1,2)(1,0,0)[12]
```

```
Q* = 15.295, df = 5, p-value = 0.009173

Model df: 5.   Total lags used: 10
```



Interpretation: While the residuals are relatively symmetric and normally distributed, the Ljung-Box test and the ACF plot suggest that the model may not have fully captured all of the autocorrelation present in the data, indicate a need to revisit the model selection.

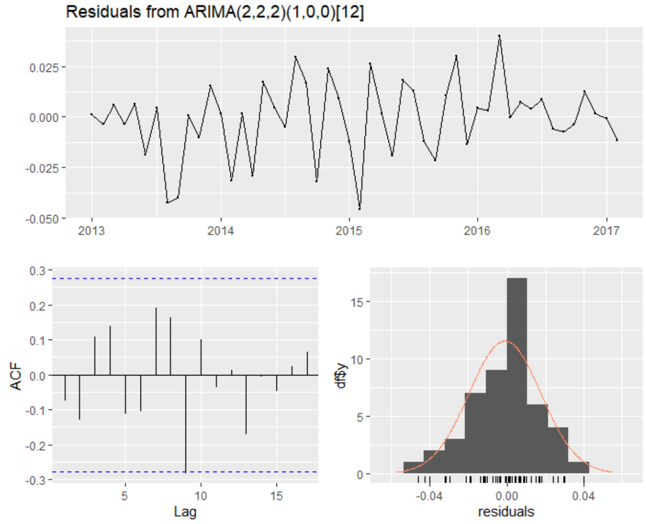
Diagnostic checks for the Squared Root Transformed ARIMA model

```
> checkresiduals(fit_sqrt_arima)

        Ljung-Box test

data:  Residuals from ARIMA(2,2,2)(1,0,0)[12]
Q* = 13.922, df = 5, p-value = 0.01612

Model df: 5.   Total lags used: 10
```

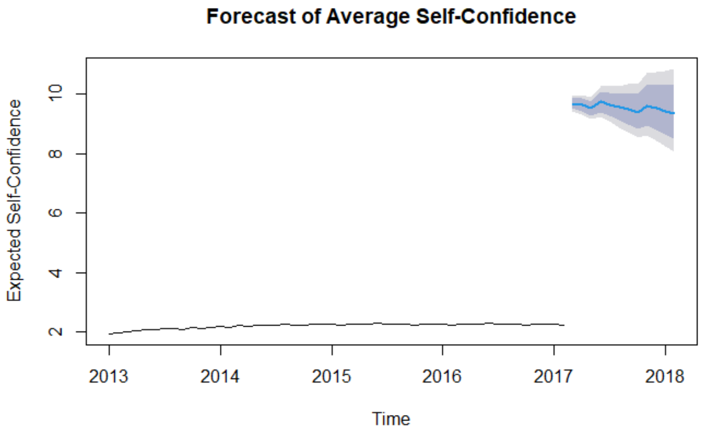


Interpretation: This model seems to have a good fit, as evidenced by the random pattern of residuals and their approximate normal distribution. However, the Ljung-Box test does suggest some remaining autocorrelation, which could indicate that further model refinement might be beneficial.

3. Make a prediction for the expected average self-confidence for the coming months.

Given that the log-transformed ARIMA model was determined to be the best fit based on both AIC and BIC criteria, we will use this model to make predictions for the expected average self-confidence in the coming months (next 12 months)

```
> # Forecast future values using the chosen model
> forecasts <- forecast(fit_log_arima, h=12)
>
> # Since we used a log transformation, we need to back-transform the pr
edictions
> forecasts$mean <- exp(forecasts$mean)
> forecasts$lower <- exp(forecasts$lower)
> forecasts$upper <- exp(forecasts$upper)
>
> # Plot the forecasts including the back-transformed predictions
> plot(forecasts, main="Forecast of Average Self-Confidence", xlab="Time
", ylab="Expected Self-Confidence")
>
> # Print the forecasted values and prediction intervals
> print(forecasts$mean)
          Jan      Feb      Mar      Apr      May      Jun      Jul
2017                   9.692767 9.637004 9.527535 9.742900 9.664945
2018 9.440787 9.361187
          Aug      Sep      Oct      Nov      Dec
2017 9.566026 9.496077 9.405660 9.607994 9.532157
2018
> print(forecasts$lower)
              80%      95%
Mar 2017 9.526974 9.440360
Apr 2017 9.426019 9.316207
May 2017 9.285117 9.159296
Jun 2017 9.402991 9.227881
Jul 2017 9.284017 9.088481
Aug 2017 9.121886 8.895181
Sep 2017 8.980303 8.718708
Oct 2017 8.838046 8.551570
Nov 2017 8.947035 8.615755
Dec 2017 8.804654 8.442290
Jan 2018 8.648311 8.256093
Feb 2018 8.493467 8.067173
> print(forecasts$upper)
               80%       95%
Mar 2017  9.861444  9.951922
Apr 2017  9.852712  9.968848
May 2017  9.776282  9.910579
Jun 2017 10.095096 10.286662
Jul 2017 10.061503 10.277972
Aug 2017 10.031792 10.287464
Sep 2017 10.041473 10.342756
Oct 2017 10.009728 10.345053
Nov 2017 10.317780 10.714504
Dec 2017 10.319772 10.762722
Jan 2018 10.305881 10.795477
Feb 2018 10.317556 10.862768
```



Interpretation:

The forecasted mean self-confidence starts high in March 2017 but shows slight variability across the months, with the general trend remaining fairly stable.

The prediction intervals widen over time, which is typical in forecasts and reflects increased uncertainty the further out the prediction goes. The 95% intervals are wider than the 80% intervals, showing a broader range of possible outcomes as uncertainty increases.

The forecasted values for early 2018 (January and February) indicate a slight decrease compared to the latter months of 2017, but they remain within the range of the previous year's predictions.

The lower and upper bounds of the prediction intervals suggest the expected variability in self-confidence. For instance, in March 2017, self-confidence is expected to be between approximately 9.44 and 9.95 at the 95% confidence level.

This forecast provides a quantitative basis for planning and decision-making regarding the expected trend in self-confidence over the next year. It should be noted that these predictions assume that past patterns will continue into the future without any new external influences.