Predictive Analytics - Exercise Sheet 1

(graded)

Prof. Dr. Benjamin Buchwitz

2024

Task 1

Make forecasts by applying the mean, naïve and drift method to the data set `tsibbledata::hh_budget` (O'Hara-Wild et al. 2022) after filtering `Australia` and choosing `Unemployment` as the dependent variable.

a) Plot the time series and describe it.

b) Split the time series data in a training and a test set and produce forecasts for the test data. Which size have you chosen for your training and test set and why?
Which of the three forecasting methods is appropriate and why? Provide empirical and theoretical arguments. Plot all values - you have forecasted by the different methods - in **one** plot!

c) After producing forecasts based on the different simple forecasting methods in Task 1 b) you get a `fable`. What does the notation `N(value1, value2)` in the column `Unemployment` mean?

d) What happens when you apply the seasonal naïve method to the data and why does it happen?

e) For the forecasting method you have chosen as the most appropriate one in Task 1 b), examine its residuals.

e.1) Which residuals do you have to consider? Give an explanation.

e.2) Plot the residuals in form of a time series, a histogram and an acf plot. What can you see?

e.3) Additionally, use the Ljung-Box test to decide whether the autocorrelations come from a white noise series or not. What are the hypotheses of the Ljung-Box test? If the autocorrelations did come from white noise, the test statistic would follow a χ^2 distribution. How many degrees of freedom would the χ^2 distribution then have?

Task 2

a) What are prediction intervals? How can you calculate them?

b) How can prediction intervals be understood in the context of probabilities? Make a sketch.

c) The standard deviation of the residuals $\hat{\sigma}$:

$\hat{\sigma} = \sqrt{\frac{1}{T-K}\sum_{t=1}^{T} e_t^2}$ (R. J. Hyndman and Athanasopoulos 2021)

is based on the sum of squared residuals $\left(e_t^2\right)$. Why is $\hat{\sigma}$ not based on the sum of residuals (e_t)?

d) How can you get the values given in Table 5.1 in Chapter 5.5 in R. J. Hyndman and Athanasopoulos (2021)? Show it exemplarily for the prediction intervals

d.1) with 95 % coverage probability and

d.2) with 50 % coverage probability.

e) After applying the simple forecasting methods to the `hh_budget` data set in Task 1, you can get the prediction intervals of your forecasts by using the function `distributional::hilo()` (O'Hara-Wild, Kay, and Hayes 2022). By default, `hilo()` returns the 80 % and 95 % prediction intervals.
What is assumed when calculating prediction intervals?
For the most appropriate forecasting method you have chosen in Task 1: Calculate the 99 % prediction intervals for your forecasts manually in R. Compare your results with the ones which you can get with `hilo()`.

Task 3

Use the Apple closing stock prices from the data set `tsibbledata::gafa_stock` (O'Hara-Wild et al. 2022) to do the following:

a) Filter the year 2016, add a daily time index to the filtered data since stock prices are not observed every day and save it as `appl_2016`. Plot `appl_2016` and describe it.

b) Split `appl_2016` in a training and test data set. Which size have you chosen for your training and test set and why?

c) Produce forecasts for the test data using the mean, naïve and drift method and plot the results together with the time series `appl_2016` in **one** plot. Which of the three forecasting methods is appropriate and why? Provide empirical and theoretical arguments.

d) For the forecasting method you have chosen as the most appropriate one in Task 3 c, examine the residuals.

d.1) Which residuals do you consider? Give an explanation.

d.2) Plot the residuals in form of a time series, a histogram and an acf plot. What can you see?

d.3) Use the Ljung-Box test to decide whether the autocorrelations come from a white noise series or not. If the autocorrelations did come from white noise, the test statistic would follow a χ^2 distribution. How many degrees of freedom would the χ^2 distribution then have?

Task 4

Make yourself familiar with the data set `fpp3::us_change` (R. Hyndman 2021). A linear regression model has been fitted to the data as you can see from the following output:

```
fit_consMR <- us_change %>%
  model(tslm = TSLM(Consumption ~ 1 + Income + Production +
                                    Unemployment + Savings))
report(fit_consMR)
```

```
## Series: Consumption
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90555 -0.15821 -0.03608  0.13618  1.15471
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.253105   0.034470   7.343 5.71e-12 ***
## Income        0.740583   0.040115  18.461  < 2e-16 ***
## Production    0.047173   0.023142   2.038   0.0429 *
## Unemployment -0.174685   0.095511  -1.829   0.0689 .
```

```
## Savings      -0.052890   0.002924 -18.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3102 on 193 degrees of freedom
## Multiple R-squared: 0.7683,  Adjusted R-squared: 0.7635
## F-statistic:   160 on 4 and 193 DF, p-value: < 2.22e-16
```

a) Write down the equation of the linear regression model and interpret the relationship between the forecast variable and the predictors.

b) What does the `1 +` mean in the code chunk of this task?

c) Compute the residual standard error that is given in the above output manually in R.

d) Show an alternative way how

d.1) you can fit a linear model to the `us_change` data set as it has been done here in this Task 4 based on `fable::TSLM()` (O'Hara-Wild, Hyndman, and Wang 2021a).

d.2) you can get the same information about the model as it has been presented here in this Task 4 with `fabletools::report(fit_consMR)` (O'Hara-Wild, Hyndman, and Wang 2021b).

Task 5

Here, the data set `tsibbledata::global_economy` (O'Hara-Wild et al. 2022) will be taken into account by focusing on Spain.

a) Plot Imports vs. Population from Spain and fit a linear model.

a.1) Write down the equation of the linear regression model by using the variables' names and interpret the relation.

a.2) What value does the multiple R^2 take and how can you interpret it?

b) Evaluate the linear regression model by plotting

b.1) the fitted values against the actual values. Add a line that goes through the origin (0,0) with a slope of 1. Describe the results and interpret them.

b.2) the residuals in form of a time series, an acf plot and a histogram. Furthermore, do a Ljung-Box test. Describe the results and interpret them.

b.3) the residuals against the predictors. Describe the results and interpret them.

b.4) the residuals against the fitted values. Describe the results and interpret them.

c) Based on the linear regression model you have fitted to the data in Task 5 a), do a one-step forecast (for 2018) for the Spanish imports. Assume that the Spanish population comprised 47,000,000 people in 2018. Plot the time series and the forecast with prediction intervals. Do you think your forecast is reliable? Why or why not?

d) Compute the adjusted R^2, CV, AIC, AICc and BIC of your fitted model from Task 5 a). Can you find more suitable (AICc) predictors to explain the imports in the Spanish economy data? Repeat the steps you did in Task 5 a) and b) for your chosen model!

Task 6

An elasticity coefficient is the ratio of the percentage change in the forecast variable (y) to the percentage change in the predictor variable (x). Mathematically, the elasticity is defined as $\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) \cdot \left(\frac{x}{y}\right)$.

Consider the log-log model: $\log y = \beta_0 + \beta_1 \log x + \epsilon$, where log is understood as the natural logarithm (ln).

a) Express y as a function of x.

b) Show that the coefficient β_1 is the elasticity coefficient. Show all your intermediate steps.

References

Hyndman, Rob. 2021. *Fpp3: Data for "Forecasting: Principles and Practice" (3rd Edition)*. https://CRAN.R-project.org/package=fpp3.

Hyndman, Rob J, and George Athanasopoulos. 2021. *Forecasting: Principles and Practice*. 3rd ed. Springer-Lehrbuch. Melbourne, Australia: OTexts.

O'Hara-Wild, Mitchell, Rob Hyndman, and Earo Wang. 2021a. *Fable: Forecasting Models for Tidy Time Series*. https://CRAN.R-project.org/package=fable.

———. 2021b. *Fabletools: Core Tools for Packages in the 'Fable' Framework*. https://CRAN.R-project.org/package=fabletools.

O'Hara-Wild, Mitchell, Rob Hyndman, Earo Wang, and Rakshitha Godahewa. 2022. *Tsibbledata: Diverse Datasets for 'Tsibble'*. https://CRAN.R-project.org/package=tsibbledata.

O'Hara-Wild, Mitchell, Matthew Kay, and Alex Hayes. 2022. *Distributional: Vectorised Probability Distributions*. https://CRAN.R-project.org/package=distributional.