Extra Credit

Zahlen Zbinden

2024-03-18

```r
# set modifiable common theme for all plots
com_theme <- function(...) {
  theme(
    plot.title=element_text(size=12, face="bold"),
    axis.title=element_text(size=8, face="bold"),
    axis.title.x=element_text(vjust=-1),
    axis.title.y=element_text(vjust=2),
    panel.grid.minor=element_blank(),
    panel.background=element_rect(fill="#FFFFED"),
    plot.background=element_rect(fill="#FFFFED")
  ) +
  theme(...)
}
```

```r
data(NileMin)
nile <- NileMin
```

```r
start_date <- as.Date("622-01-01")
end_date <- as.Date("1284-12-31")
```

```r
length(seq(start_date, end_date, by = "year"))
```

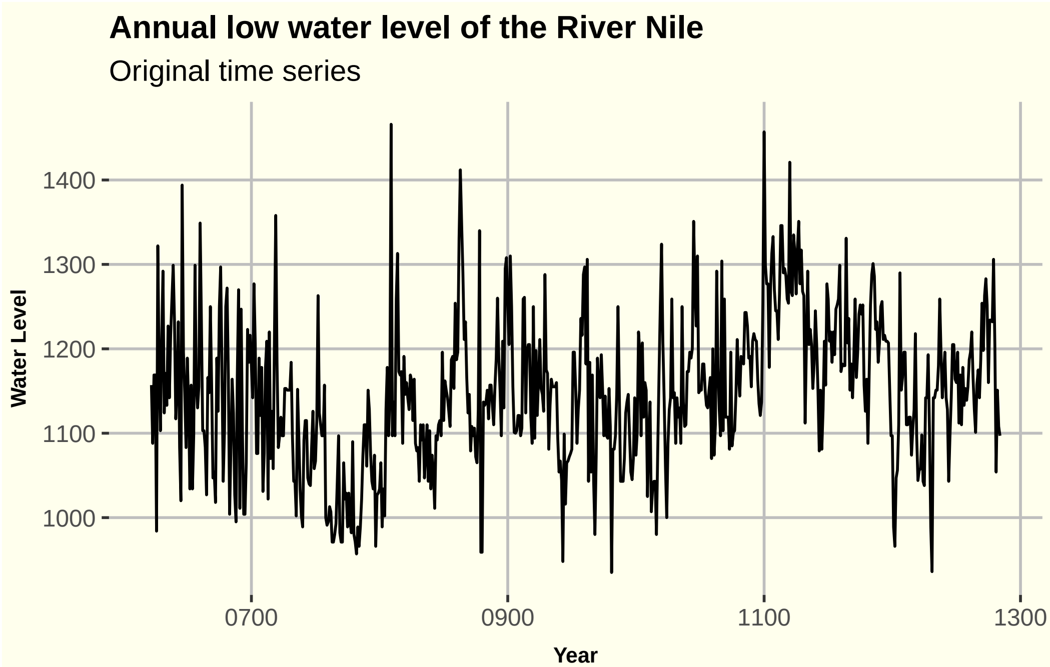
```
[1] 663
```

a) Display the time series plots of the original data and the first difference of the original data.

```r
# plot the original time series

nile |>
    as_tibble() |>
    mutate(
        date=seq(
            start_date,
            end_date,
            by="year"
        )
    ) |>
    ggplot(aes(x=date, y=x)) +
        geom_line() +
        labs(
          title="Annual low water level of the River Nile",
          subtitle="Original time series",
          x="Year",
          y="Water Level"
        ) +
        com_theme(
          panel.grid.major=element_line(color="grey")
        )
```

```
Don't know how to automatically pick scale for object of type <ts>. Defaulting
to continuous.
```



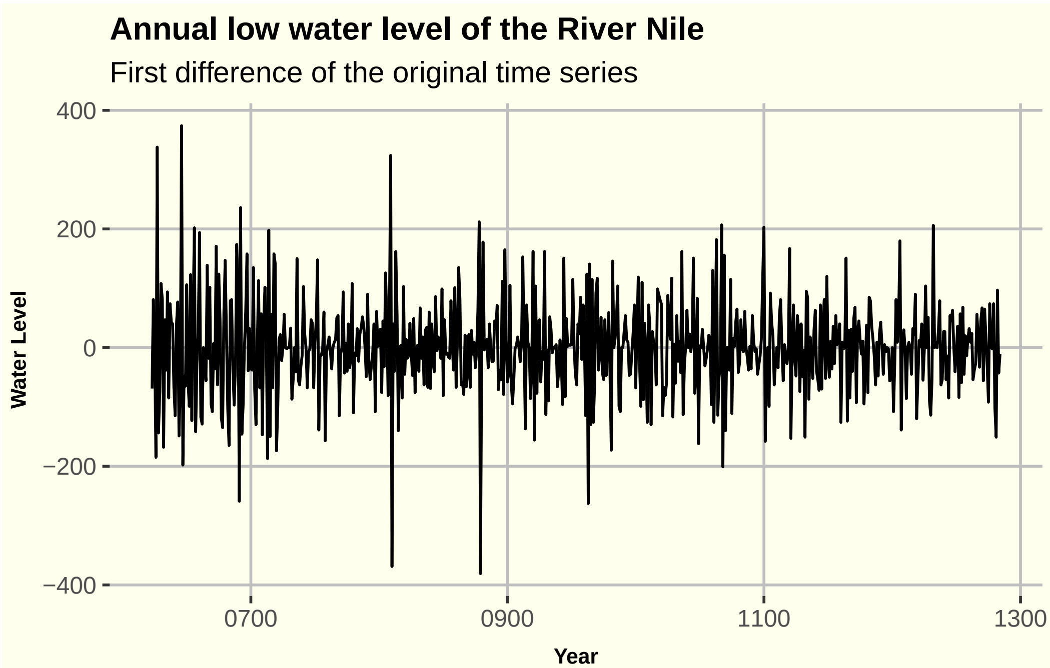
b)

```
# first difference of the original time series

nile |>
    diff() |>
    as_tibble() |>
    mutate(
        date=seq(
            start_date + years(1),
            end_date,
            by="year"
        )
    ) |>
    ggplot(aes(x=date, y=x)) +
        geom_line() +
        labs(
          title="Annual low water level of the River Nile",
          subtitle="First difference of the original time series",
          x="Year",
          y="Water Level"
        ) +
```

```
        com_theme(
          panel.grid.major=element_line(color="grey")
        )
```

```
Don't know how to automatically pick scale for object of type <ts>. Defaulting
to continuous.
```



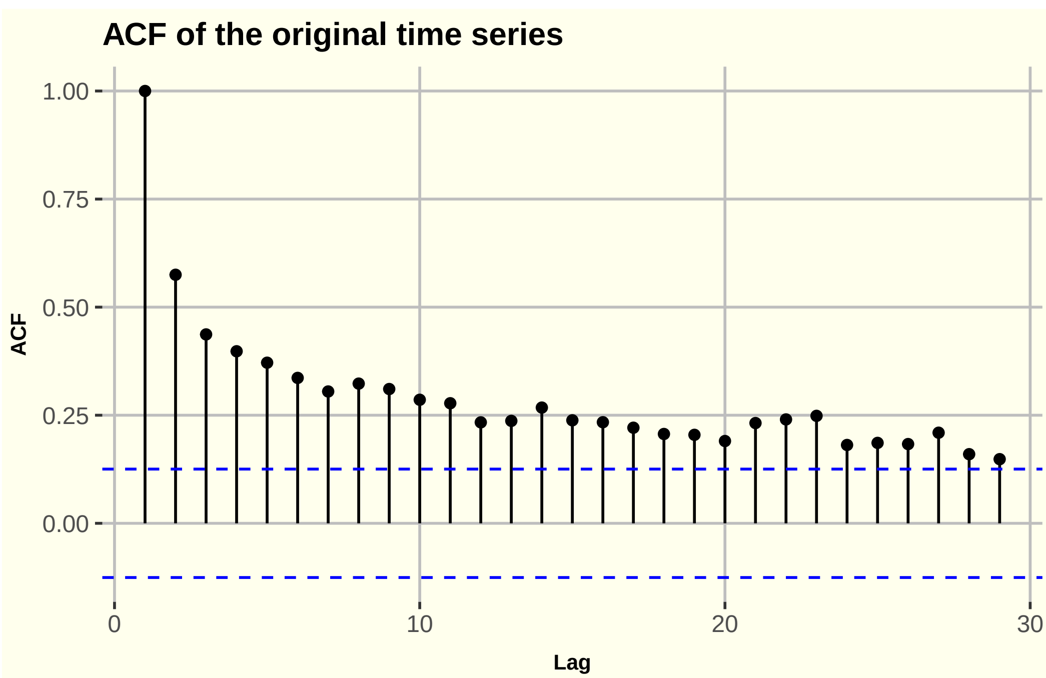
b) Calculate the acf and pacf for the original time series and the first differenced time series.

ACF and PACF of the original time series

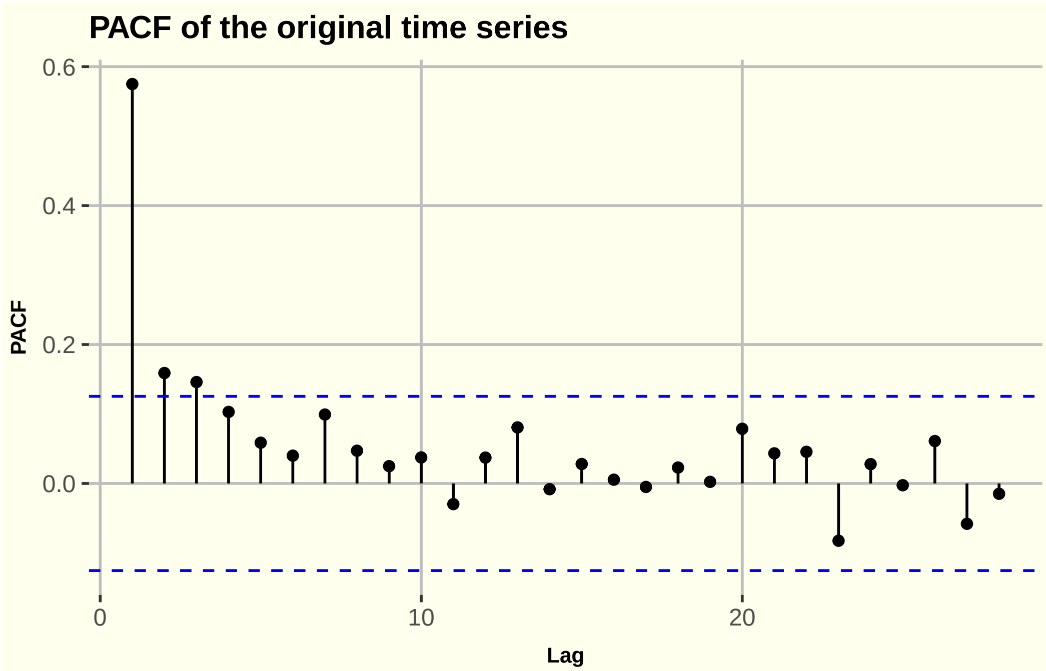
```
# acf and pacf for the original time series
acf_og <- acf(nile, plot=FALSE)
pacf_og <- pacf(nile, plot=FALSE)
ci <- qnorm((1 - .05) / 2) / sqrt(length(acf_og$n.used))
```

```
acf_og$acf |>
  as_tibble() |>
  mutate(
    lag=1:length(acf_og$acf)
```

```
) |>
ggplot(aes(x=lag, y=V1)) +
  geom_segment(
    aes(
      x=lag,
      xend=lag,
      y=0,
      yend=V1
    )
  ) +
  geom_point() +
  geom_hline(yintercept=2 * ci, color="blue", linetype="dashed") +
  geom_hline(yintercept=-2 * ci, color="blue", linetype="dashed") +
  labs(
    title="ACF of the original time series",
    x="Lag",
    y="ACF"
  ) +
  com_theme(
    panel.grid.major=element_line(color="grey")
  )
```



```
pacf_og$acf |>
  as_tibble() |>
  mutate(
    lag=1:length(pacf_og$acf)
  ) |>
  ggplot(aes(x=lag, y=V1)) +
    geom_segment(
      aes(
        x=lag,
        xend=lag,
        y=0,
        yend=V1
      )
    ) +
    geom_point() +
    geom_hline(yintercept=2 * ci, color="blue", linetype="dashed") +
    geom_hline(yintercept=-2 * ci, color="blue", linetype="dashed") +
    labs(
      title="PACF of the original time series",
      x="Lag",
      y="PACF"
    ) +
    com_theme(
      panel.grid.major=element_line(color="grey")
    )
```

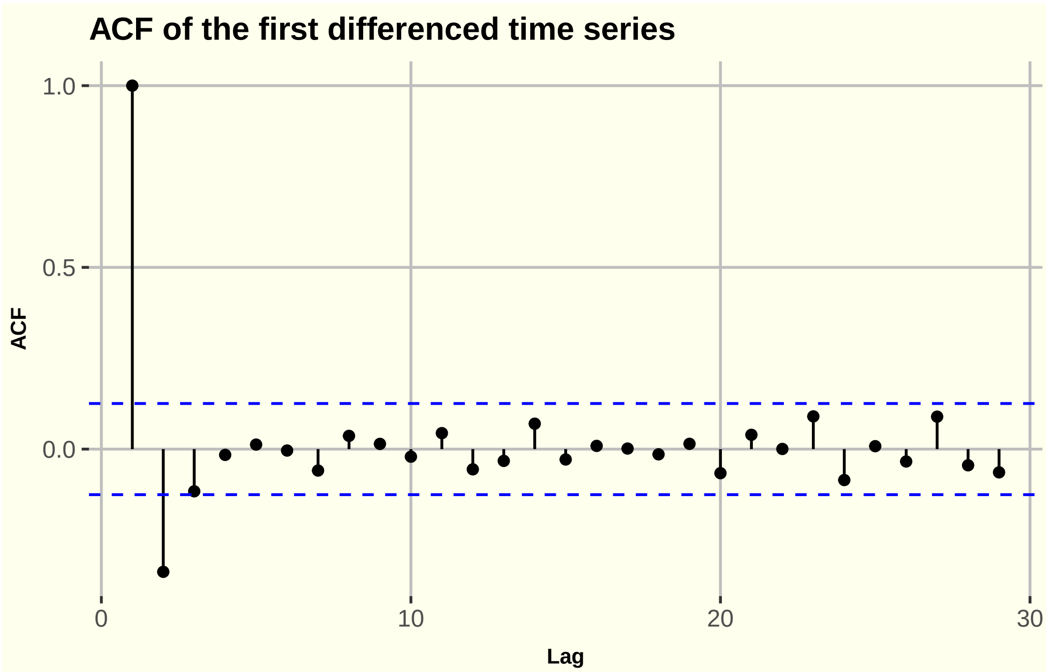


ACF and PACF of the first differenced time series

```
acf_dif <- acf(diff(nile), plot=FALSE)
pacf_dif <- pacf(diff(nile), plot=FALSE)
ci <- qnorm((1 - .05) / 2) / sqrt(length(acf_dif$n.used))
```

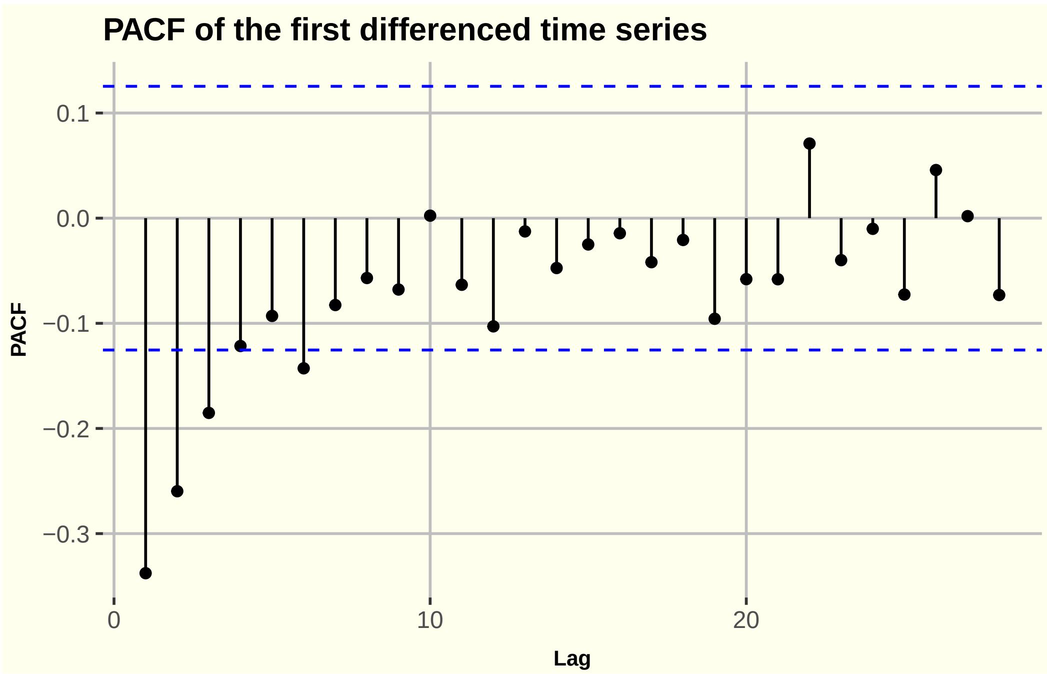
```
acf_dif$acf |>
  as_tibble() |>
  mutate(
    lag=1:length(acf_dif$acf)
  ) |>
  ggplot(aes(x=lag, y=V1)) +
    geom_segment(
      aes(
        x=lag,
        xend=lag,
        y=0,
        yend=V1
      )
    ) +
```

```
    geom_point() +
    geom_hline(yintercept=2 * ci, color="blue", linetype="dashed") +
    geom_hline(yintercept=-2 * ci, color="blue", linetype="dashed") +
    labs(
      title="ACF of the first differenced time series",
      x="Lag",
      y="ACF"
    ) +
    com_theme(
      panel.grid.major=element_line(color="grey")
    )
```




```
pacf_dif$acf |>
  as_tibble() |>
  mutate(
    lag=1:length(pacf_dif$acf)
  ) |>
  ggplot(aes(x=lag, y=V1)) +
    geom_segment(
      aes(
        x=lag,
```

```
      xend=lag,
      y=0,
      yend=V1
    )
  ) +
  geom_point() +
  geom_hline(yintercept=2 * ci, color="blue", linetype="dashed") +
  geom_hline(yintercept=-2 * ci, color="blue", linetype="dashed") +
  labs(
    title="PACF of the first differenced time series",
    x="Lag",
    y="PACF"
  ) +
  com_theme(
    panel.grid.major=element_line(color="grey")
  )
```



c) Fit an ARIMA model to the original series, what model do you choose and why?

We can see from the first differencing of the series that we have data that is not correlated anymore, this means that we will want to choose an arima model with d=1. We can see from the ACF that there is one significant lab at 1, and from the PACF there are significant lags at 1,2, and 3. This means that we could choose, p=3, or q=1. Or make a slight deviation from those numbers, 1 in either direction. Let us also check what the auto.arima function suggests as a model.

```
auto.arima(nile)
```

```
Series: nile
ARIMA(3,1,1)

Coefficients:
         ar1     ar2     ar3      ma1
      0.3715  0.0306  0.0639  -0.9404
s.e.  0.0449  0.0436  0.0428   0.0219

sigma^2 = 4946:  log likelihood = -3753.57
AIC=7517.14   AICc=7517.23   BIC=7539.61
```

The auto arima model that is suggested is a ARIMA(3,1,1) model which would follow along with some of the thoughts that we had earlier. I will go with the ARIMA(3,1,1) model to continue our analysis with.

d) Investigate diagnostics for this model, including the AFC and normality of the residuals.

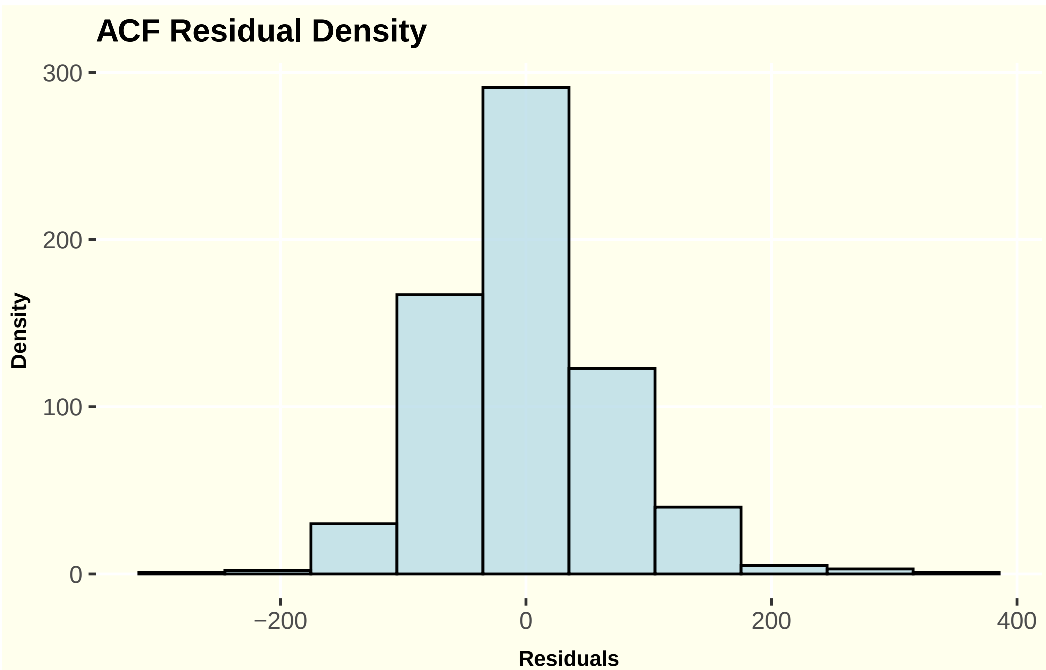
```
fit_311 <- Arima(nile, order=c(3,1,1))
```

```
res_311 <- residuals(fit_311)
res_acf <- acf(res_311, plot=FALSE)
res_pacf <- pacf(res_311, plot=FALSE)
res_ci <- qnorm((1 - .05) / 2) / sqrt(length(res_acf$n.used))
```

```
# histogram of the residuals to check for normailty assumption
res_311 |>
  as_tibble() |>
  ggplot(aes(x=x)) +
    geom_histogram(
      bins=10,
      fill="lightblue",
      color="black",
      alpha=0.7
    ) +
    labs(
      title="ACF Residual Density",
```

```
        x="Residuals",
        y="Density"
      ) +
      com_theme()
```

```
Don't know how to automatically pick scale for object of type <ts>. Defaulting
to continuous.
```

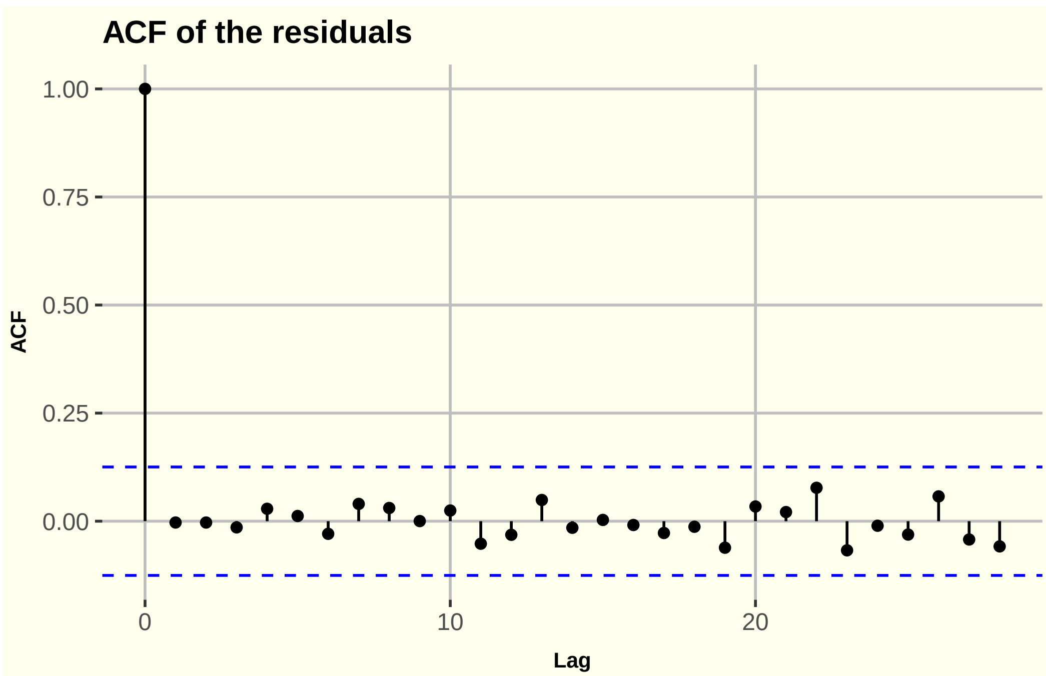


The histogram of the residual density of the ACF doesn't point us to the residuals not being evenly distributed as we can see that they follow a rough normal distribution, which is what we want when investigating the residual density.

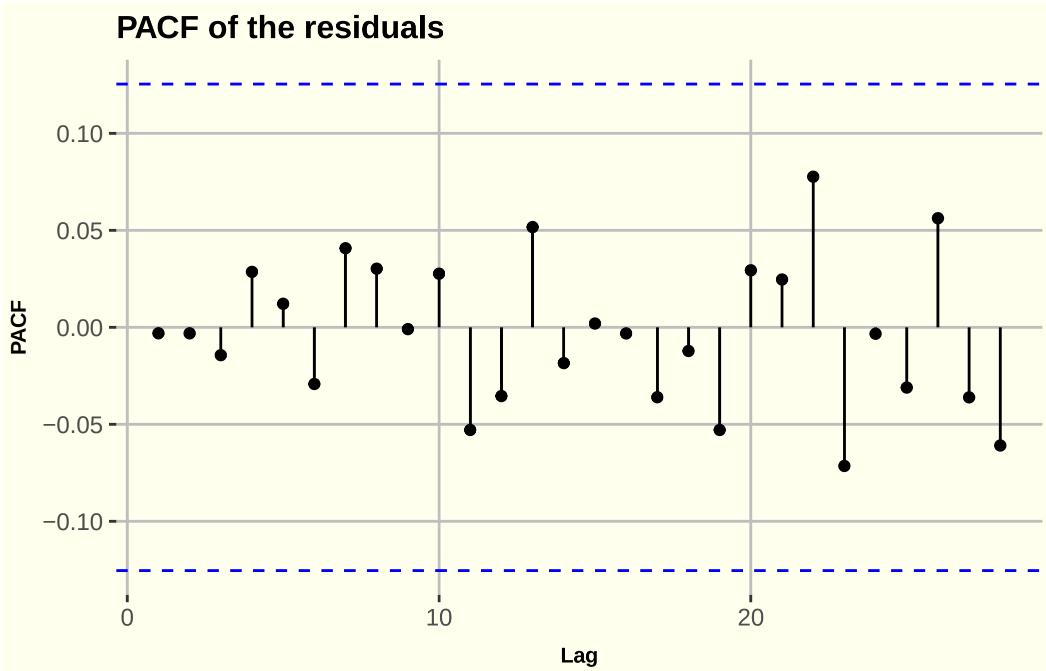
```
lag_end = length(res_acf$acf) - 1
```

```
res_acf$acf |>
  as_tibble() |>
  mutate(
    lag=0:lag_end
  ) |>
  ggplot(aes(x=lag, y=V1)) +
```

```
  geom_segment(
    aes(
      x=lag,
      xend=lag,
      y=0,
      yend=V1
    )
  ) +
  geom_point() +
  geom_hline(yintercept=2 * res_ci, color="blue", linetype="dashed") +
  geom_hline(yintercept=-2 * res_ci, color="blue", linetype="dashed") +
  labs(
    title="ACF of the residuals",
    x="Lag",
    y="ACF"
  ) +
  com_theme(
    panel.grid.major=element_line(color="grey")
  )
```



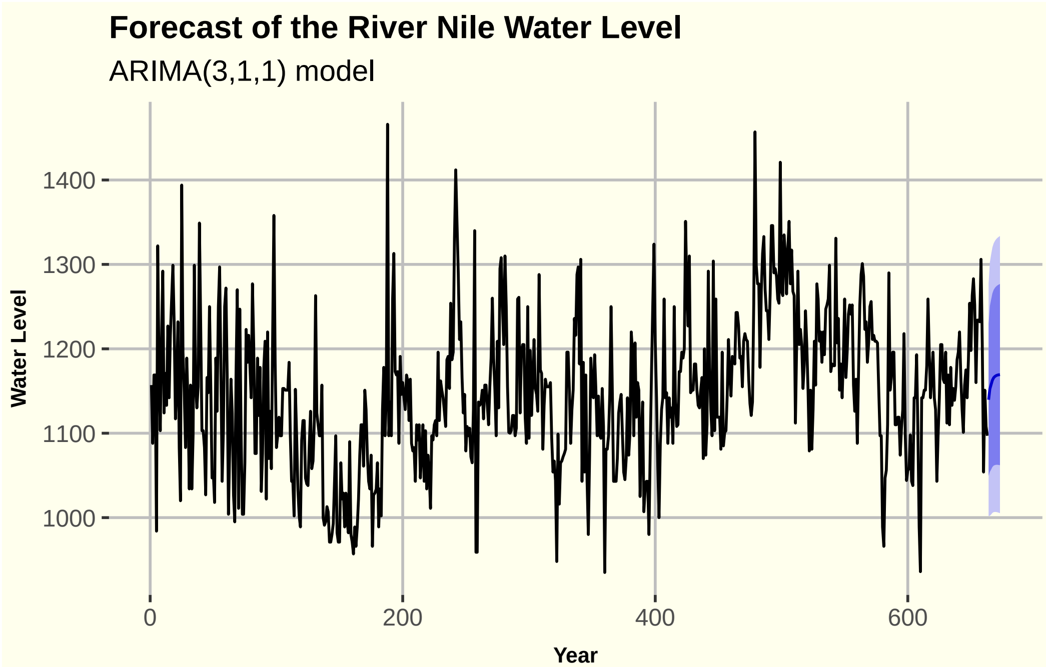
```r
res_pacf$acf |>
  as_tibble() |>
  mutate(
    lag=1:lag_end
  ) |>
  ggplot(aes(x=lag, y=V1)) +
    geom_segment(
      aes(
        x=lag,
        xend=lag,
        y=0,
        yend=V1
      )
    ) +
    geom_point() +
    geom_hline(yintercept=2 * res_ci, color="blue", linetype="dashed") +
    geom_hline(yintercept=-2 * res_ci, color="blue", linetype="dashed") +
    labs(
      title="PACF of the residuals",
      x="Lag",
      y="PACF"
    ) +
    com_theme(
      panel.grid.major=element_line(color="grey")
    )
```



We can see from both the ACf and the PACF of the residuals that there is no correlation in the lagged values. This leads us to conclude that the residuals are indeed normally distributed, and we can continue with out analysis and prediction, as this model does appear to be a satisfactory fit for out data.

e) produce forecasts for this series with a lead time of ten years. Be sure to include forecast limits.

```
autoplot(forecast(fit_311, h=10)) +
  labs(
    title="Forecast of the River Nile Water Level",
    subtitle="ARIMA(3,1,1) model",
    x="Year",
    y="Water Level"
  ) +
  com_theme(
    panel.grid.major=element_line(color="grey")
  )
```


Forecast of the River Nile Water Level
ARIMA(3,1,1) model
1400
1300
1200
1100
1000
Water Level
0
200
400
600
Year