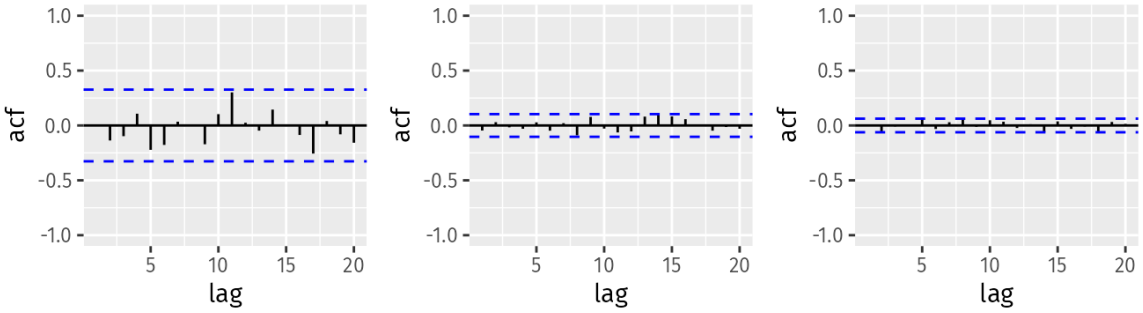
DATA 624 - Homework 6

Richie Rivera

Question 9.1

1. Explain the differences among these figures. Do they all indicate that the data are white noise?



Each of these graphs show the same y scale. We can use that to notice that the spread is smaller in each subsequent chart. The blue lines are fixed values where the autocorrelations would be statistically significantly different from zero (given by $\pm\frac{1.96}{\sqrt{(T)}}$ for 95% where T is the length of the timeseries). These charts indicate that this data is white noise as all of the ACFs values lie between those white lines.

2. Why are the critical values at different distances from the mean of zero? Why are the autocorrelations different in each figure when they each refer to white noise?

From the equation in the question before ($\pm\frac{1.96}{\sqrt{(T)}}$) we can see that there is a negative correlation with the magnitude of the critical value and the length of the timeseries. Therefore, as more time is provided the critical value decreases.

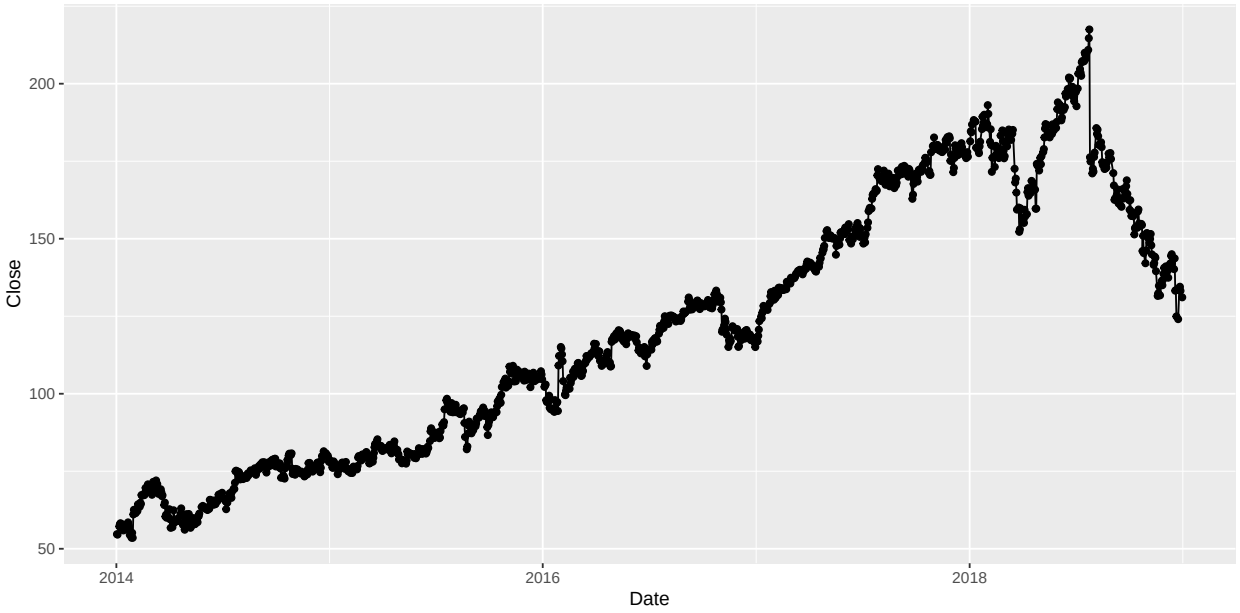
Question 9.2

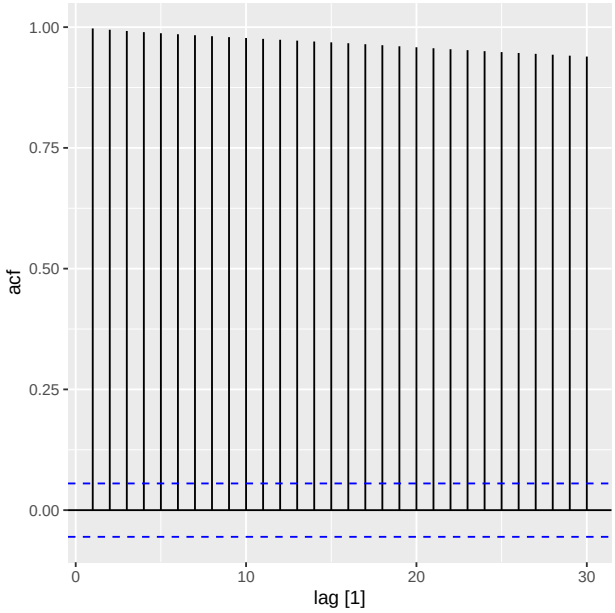
A classic example of a non-stationary series are stock prices. Plot the daily closing prices for Amazon stock (contained in gafa_stock), along with the ACF and PACF. Explain how each plot shows that the series is non-stationary and should be differenced.

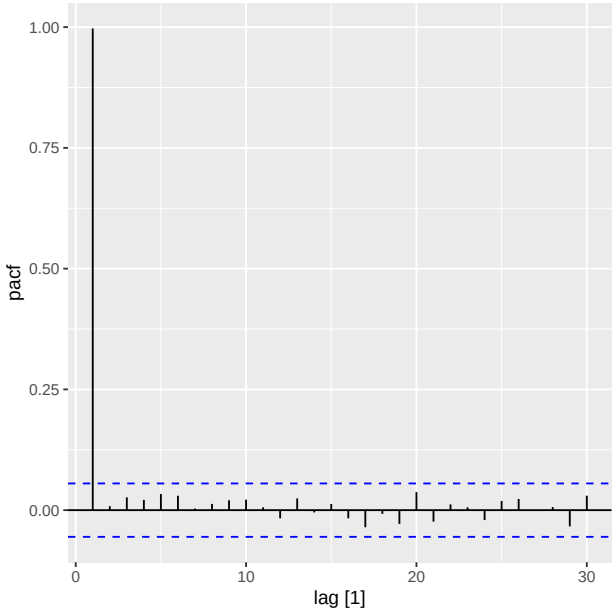
```
## # A tsibble: 6 x 8 [!]
## # Key:       Symbol [1]
##   Symbol Date        Open  High   Low Close Adj_Close   Volume
##   <chr>  <date>     <dbl> <dbl> <dbl> <dbl>     <dbl>    <dbl>
## 1 FB     2014-01-02  54.8  55.2  54.2  54.7      54.7 43195500
## 2 FB     2014-01-03  55.0  55.7  54.5  54.6      54.6 38246200
```

```
## 3 FB     2014-01-06  54.4  57.3  54.0  57.2      57.2 68852600
## 4 FB     2014-01-07  57.7  58.5  57.2  57.9      57.9 77207400
## 5 FB     2014-01-08  57.6  58.4  57.2  58.2      58.2 56682400
## 6 FB     2014-01-09  58.7  59.0  56.7  57.2      57.2 92253300
```

```
## Warning: Provided data has an irregular interval, results should be treated with caution. Computing A
## Provided data has an irregular interval, results should be treated with caution. Computing ACF by obs
```








```
## # A tibble: 1 x 3
##   Symbol kpss_stat kpss_pvalue
##   <chr>      <dbl>       <dbl>
## 1 FB          14.8        0.01
```

1. The timeseries pretty obviously has a trend which is evidence that the data is non-stationary and should be differenced.
2. From the above ACF plot, we can see that the graph slowly decays, which allows us to conclude that this dataset is non-stationary and should be differenced.
3. From the above PACF plot, we can see that there is only a strong correlation with the first lag and almost none with subsequent lags. This is an indication that the data is non-stationary and should be differenced.
4. Using the unit root test, we get a `kpss_pvalue` of .01, which could be a value much lower than .01 and is evidence that the data is non-stationary.

Question 9.3

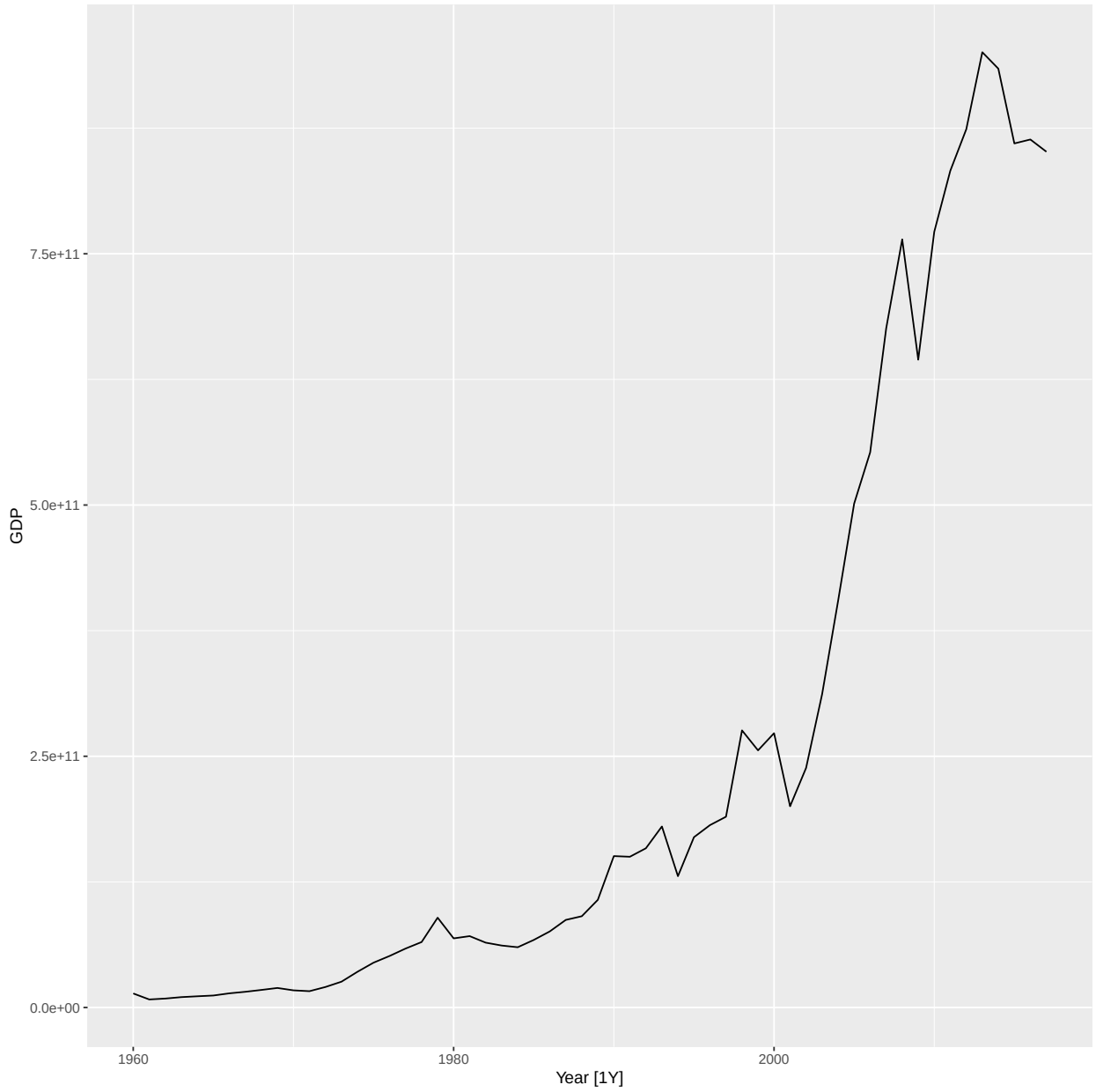
For the following series, find an appropriate Box-Cox transformation and order of differencing in order to obtain stationary data.

1. For the appropriate Box-Cox transformation, we'll find the lambda and then use the chart here to find the appropriate transformation.
2. In order to find the order of differencing, I'll use the unitroot_kpss feature.

1. Turkish GDP from `global_economy`.

```
## # A tsibble: 6 x 9 [1Y]
## # Key:       Country [1]
##   Country     Code   Year         GDP Growth   CPI Imports Exports Population
##   <fct>       <fct> <dbl>       <dbl>  <dbl> <dbl>   <dbl>   <dbl>      <dbl>
## 1 Afghanistan AFG    1960  537777811.     NA    NA    7.02    4.13    8996351
## 2 Afghanistan AFG    1961  548888896.     NA    NA    8.10    4.45    9166764
## 3 Afghanistan AFG    1962  546666678.     NA    NA    9.35    4.88    9345868
## 4 Afghanistan AFG    1963  751111191.     NA    NA   16.9     9.17    9533954
## 5 Afghanistan AFG    1964  800000044.     NA    NA   18.1     8.89    9731361
## 6 Afghanistan AFG    1965 1006666638.     NA    NA   21.4    11.3     9938414
```

```
## Plot variable not specified, automatically selected `.vars = GDP`
```

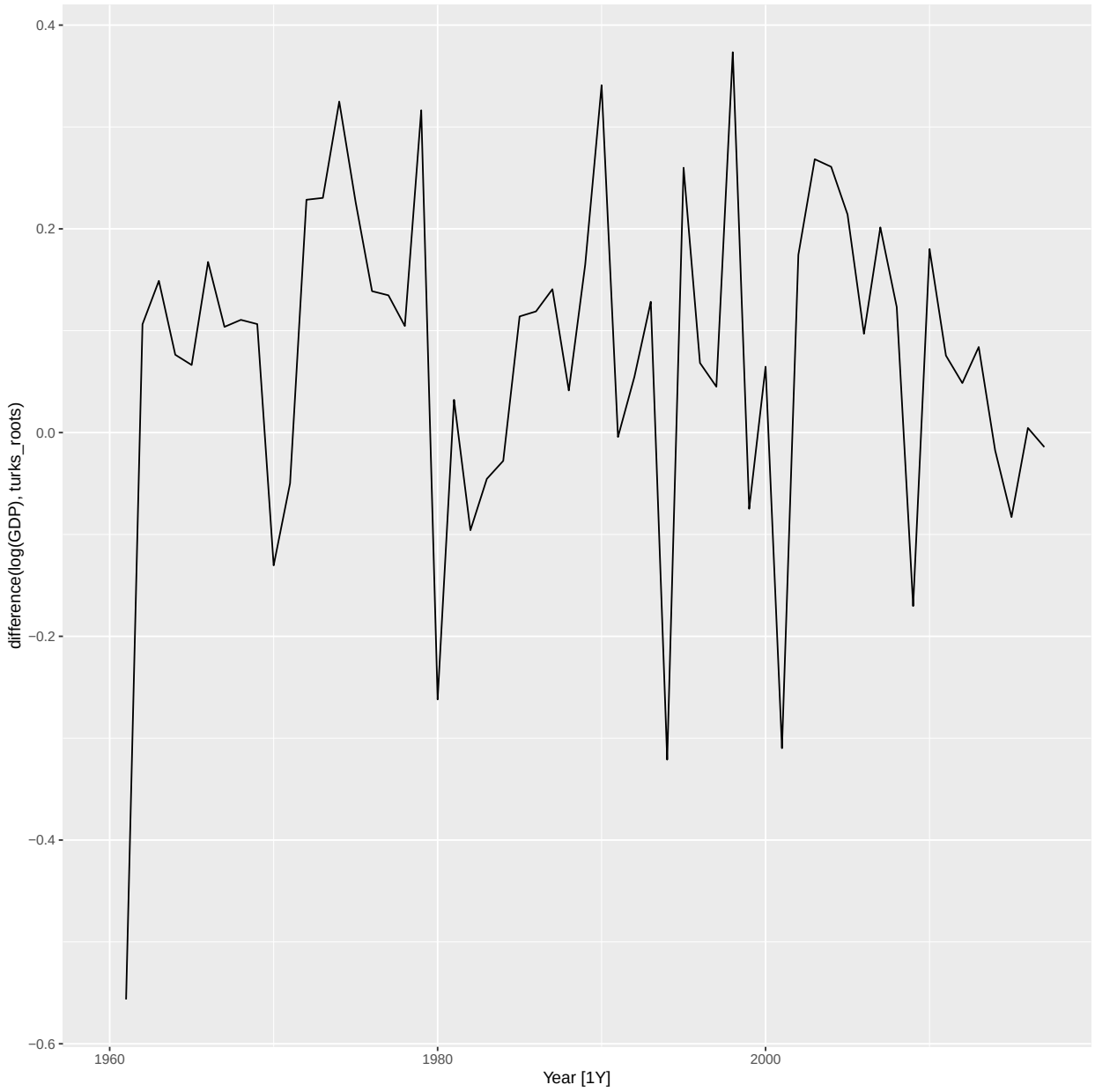



```
## [1] 0.1572187
```

This timeseries has an increasing trend. With a lambda of 0.16, we can use the $log(y)$ operation:

The kpss test recommends that we use a difference of 1.

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```

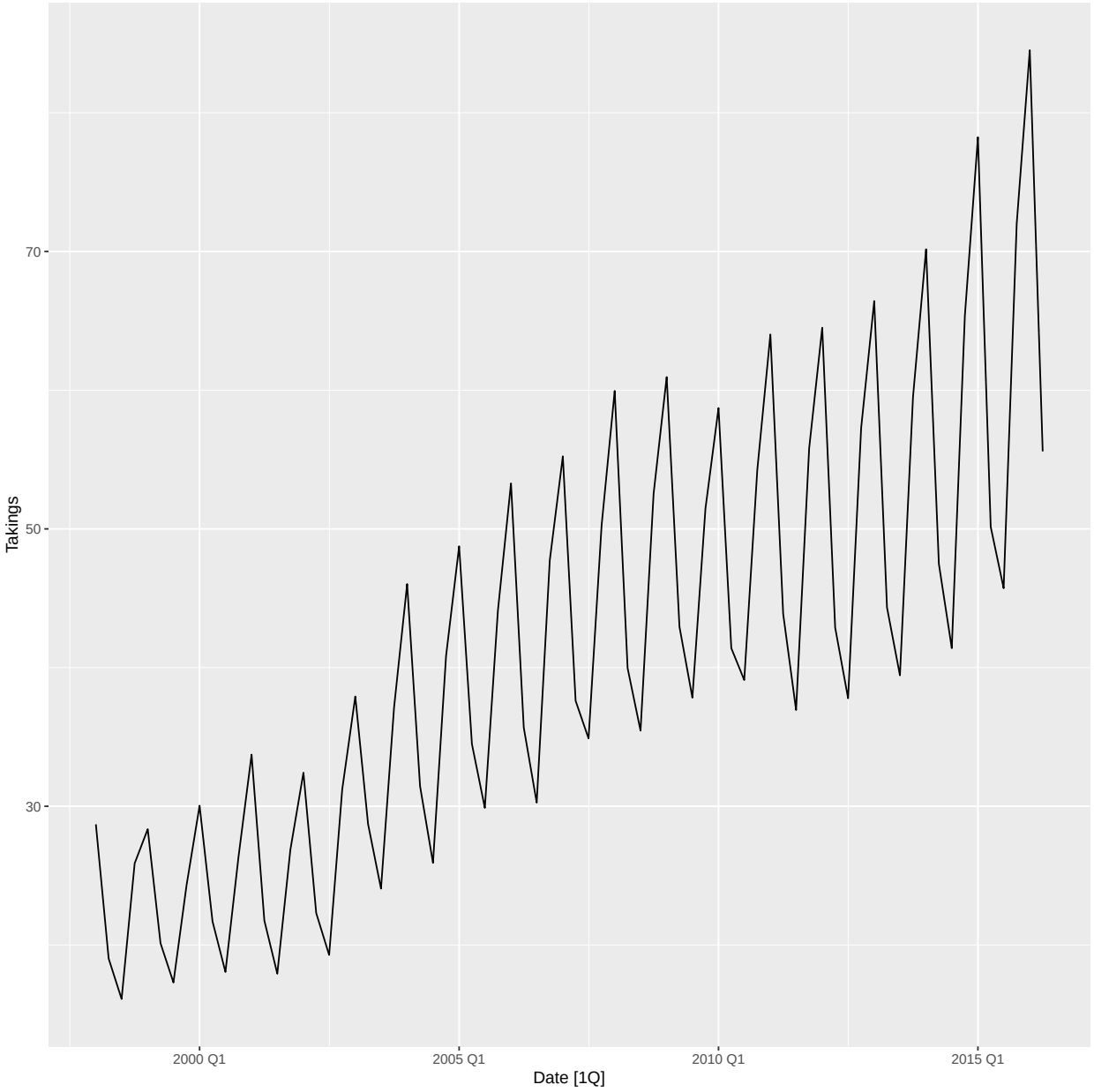


The above plot seems fairly stationary now that it's been transformed and differenced.

2. Accommodation takings in the state of Tasmania from `aus_accommodation`.

```
## # A tsibble: 6 x 5 [1Q]
## # Key:       State [1]
##      Date State                          Takings Occupancy   CPI
##     <qtr> <chr>                            <dbl>     <dbl> <dbl>
## 1 1998 Q1 Australian Capital Territory    24.3        65  67
## 2 1998 Q2 Australian Capital Territory    22.3        59  67.4
## 3 1998 Q3 Australian Capital Territory    22.5        58  67.5
## 4 1998 Q4 Australian Capital Territory    24.4        59  67.8
## 5 1999 Q1 Australian Capital Territory    23.7        58  67.8
## 6 1999 Q2 Australian Capital Territory    25.4        61  68.1
```

```
## Plot variable not specified, automatically selected `.vars = Takings`
```

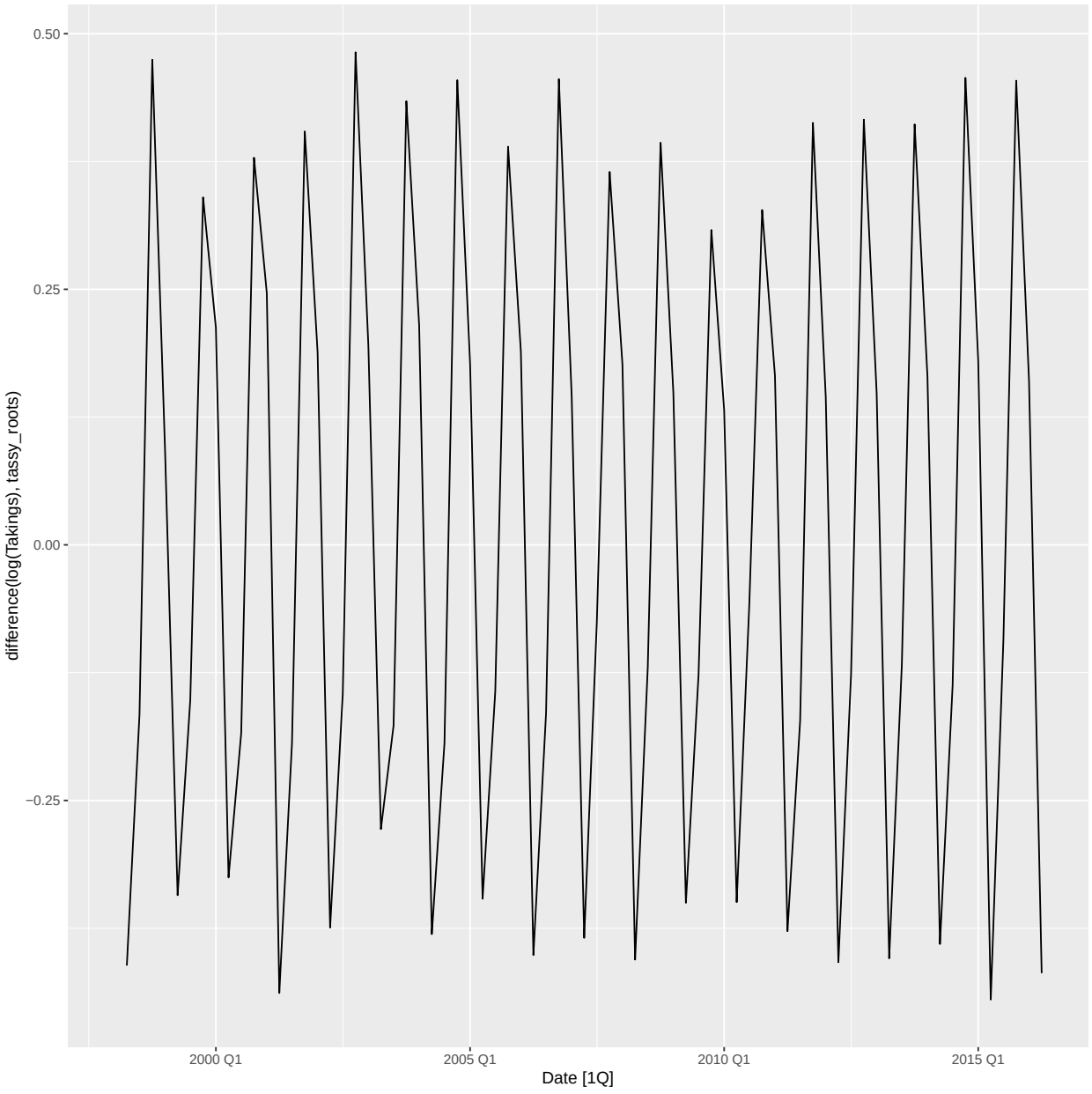



```
## [1] 0.001819643
```

This data is seasonal and has trend. With a lambda of 0, we can use the $log(y)$ operation:

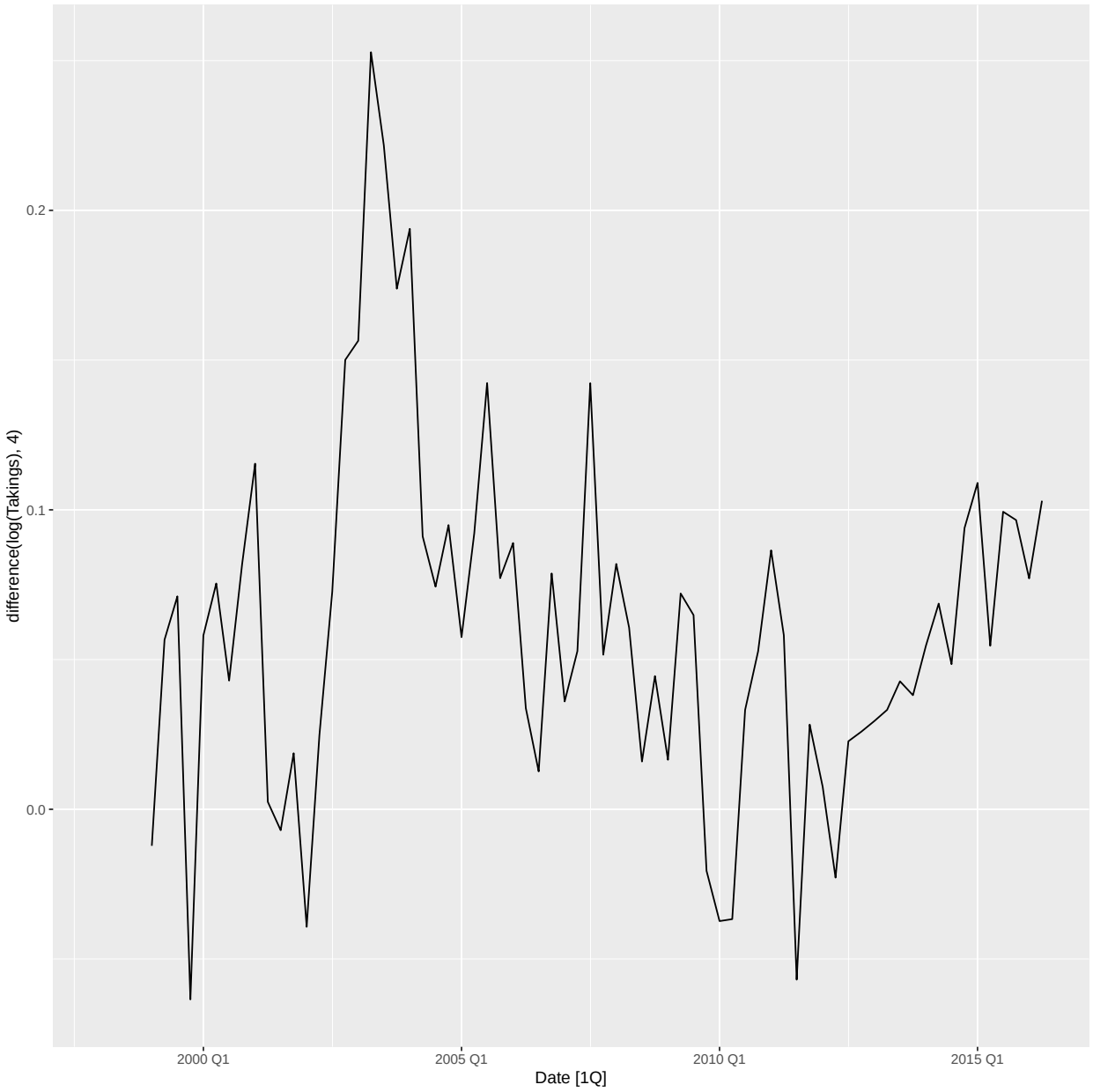
The kpss test recommends that we use a difference of 1.

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```



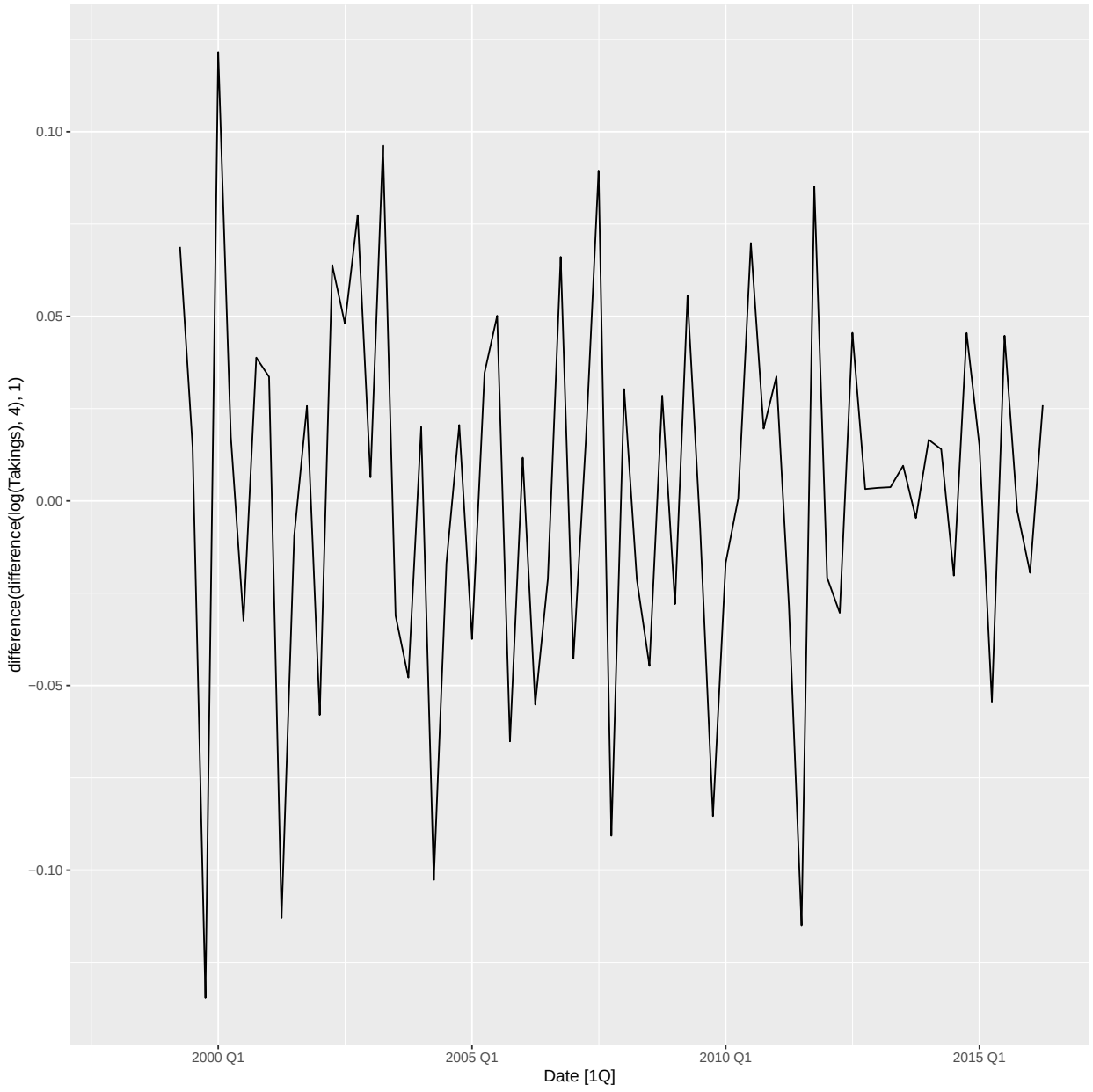
The above plot seems odd as it seems to oscillate around 0 like as sinusoidal wave. I'll be taking a difference of the seasonality duration (4 quarters):

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



After doing a difference of 4, the data doesn't seem non-stationary but it did have a different shape than a sinusoidal chart. I'll be doing a first level difference on that 4th level difference below:

```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```

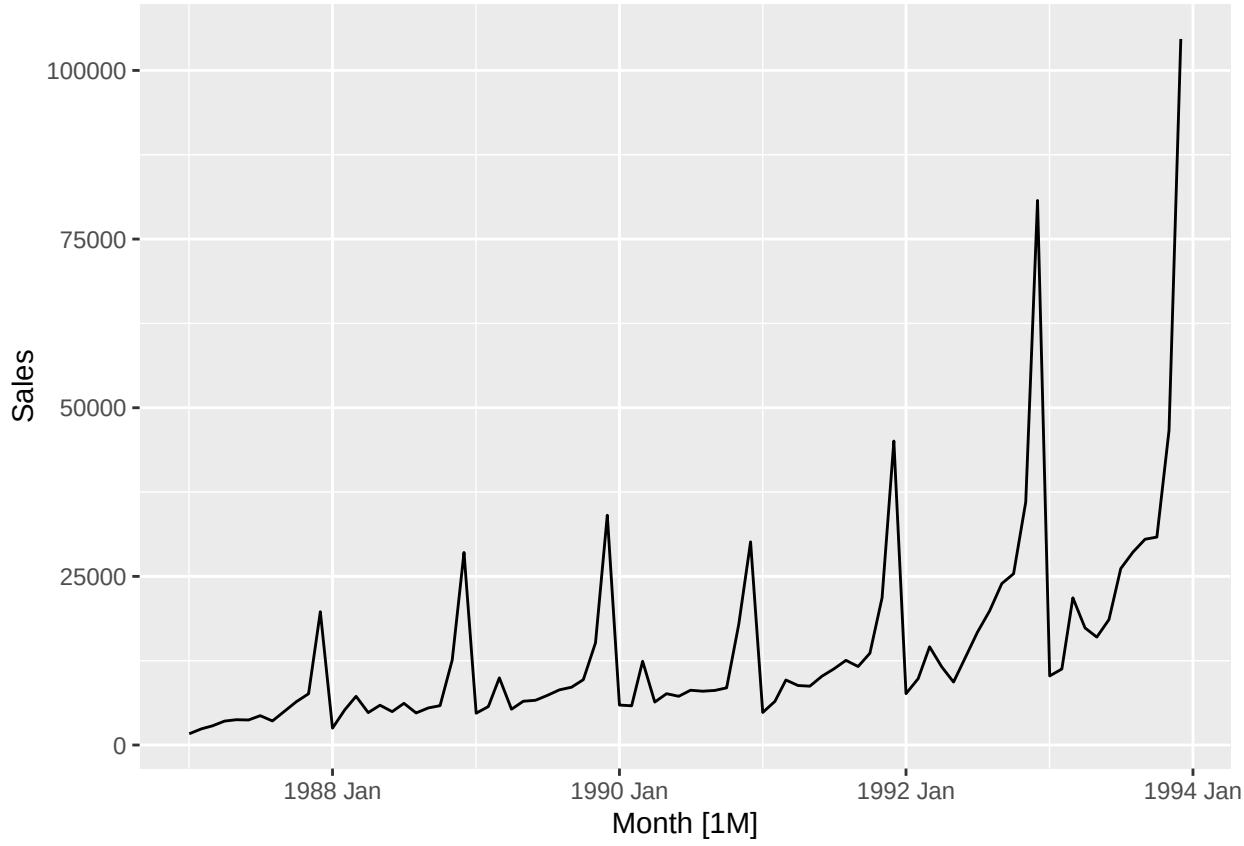


From here we see something that seems much more stationary.

3. Monthly sales from `souvenirs`.

```
## # A tsibble: 6 x 2 [1M]
##      Month Sales
##      <mth> <dbl>
## 1 1987 Jan 1665.
## 2 1987 Feb 2398.
## 3 1987 Mar 2841.
## 4 1987 Apr 3547.
## 5 1987 May 3753.
## 6 1987 Jun 3715.
```

```
## Plot variable not specified, automatically selected `.vars = Sales`
```

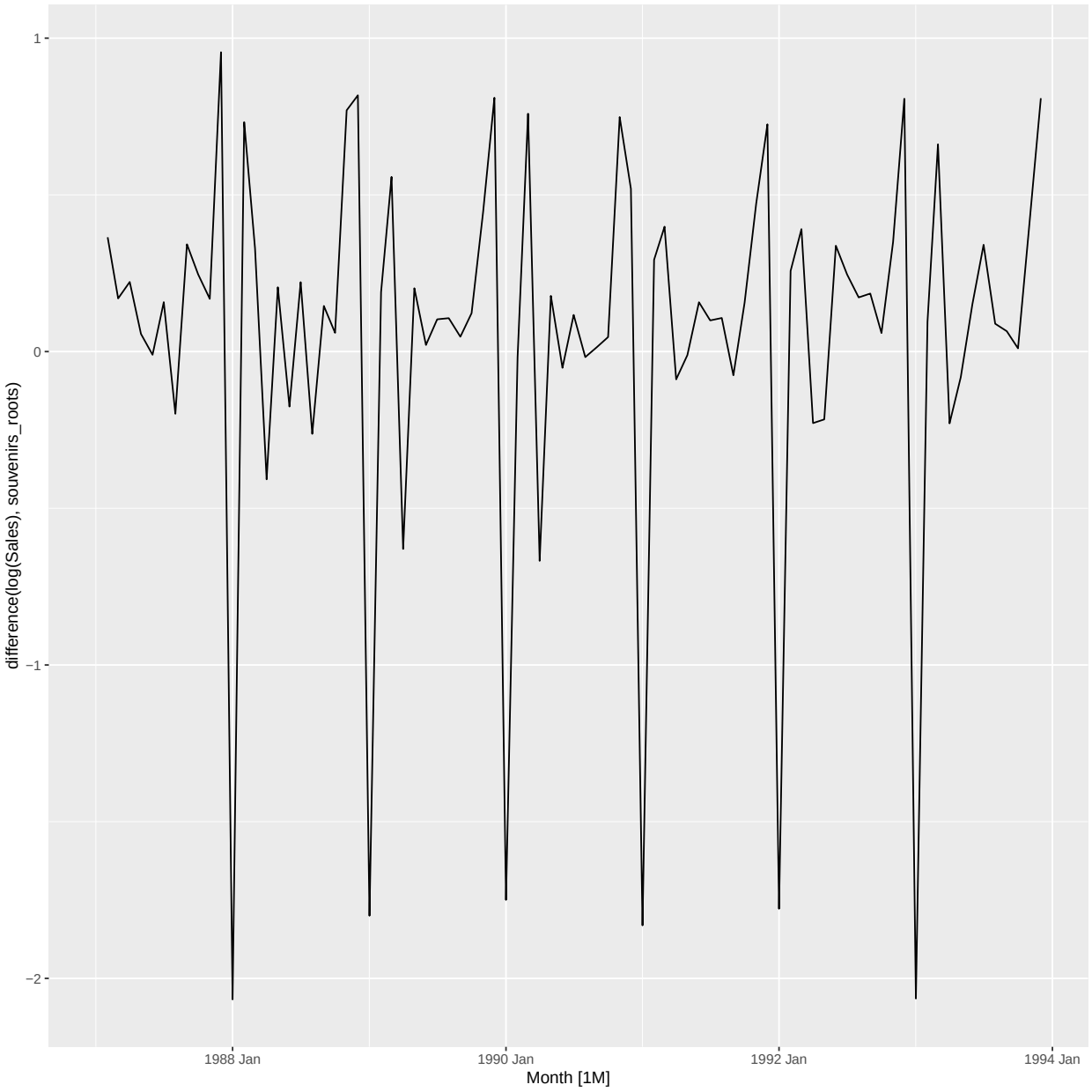



```
## [1] 0.002118221
```

This timeseries has an increasing as well as annual seasonality. With a λ of 0, we can use the $log(y)$ operation:

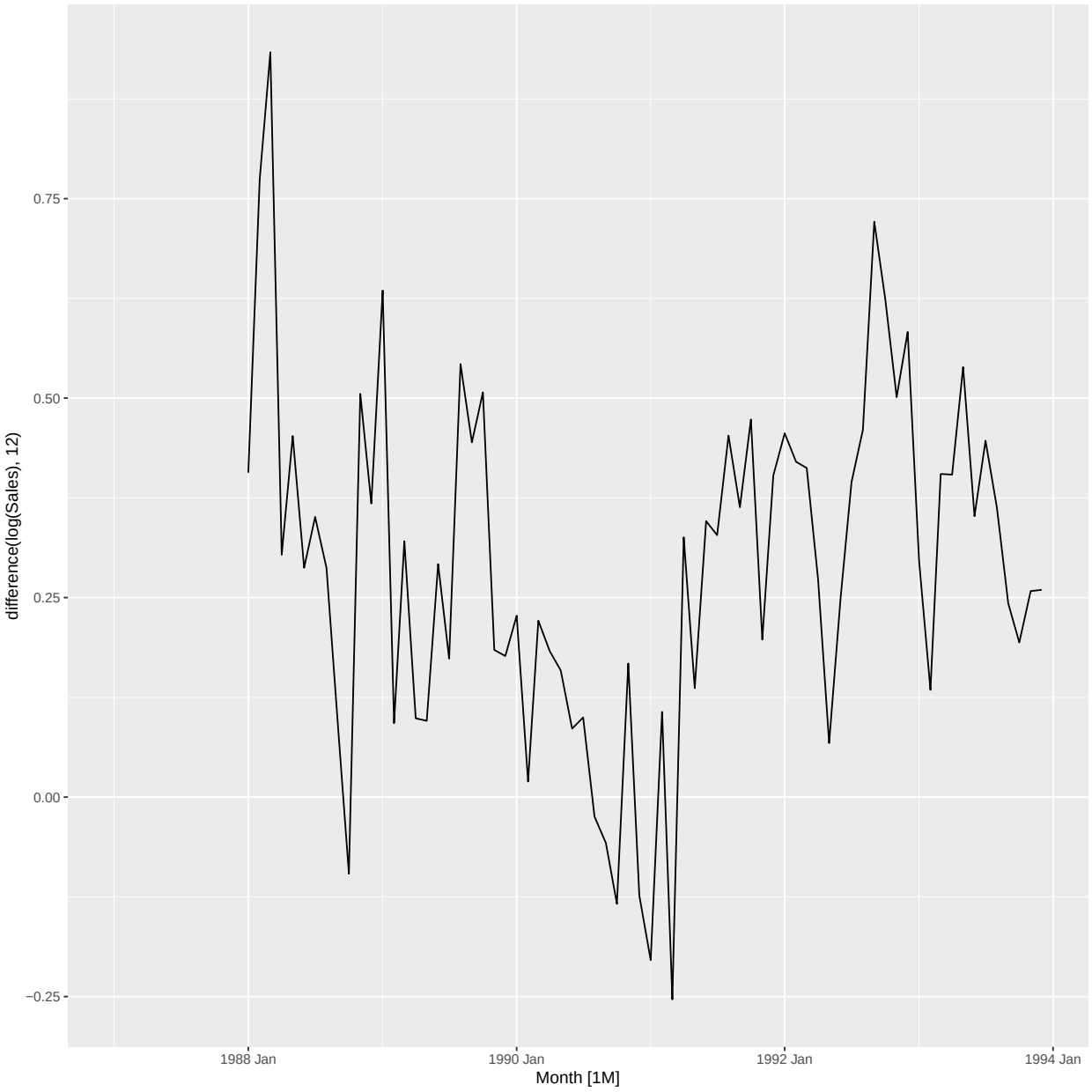
The kpss test recommends that we use a difference of 1.

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```



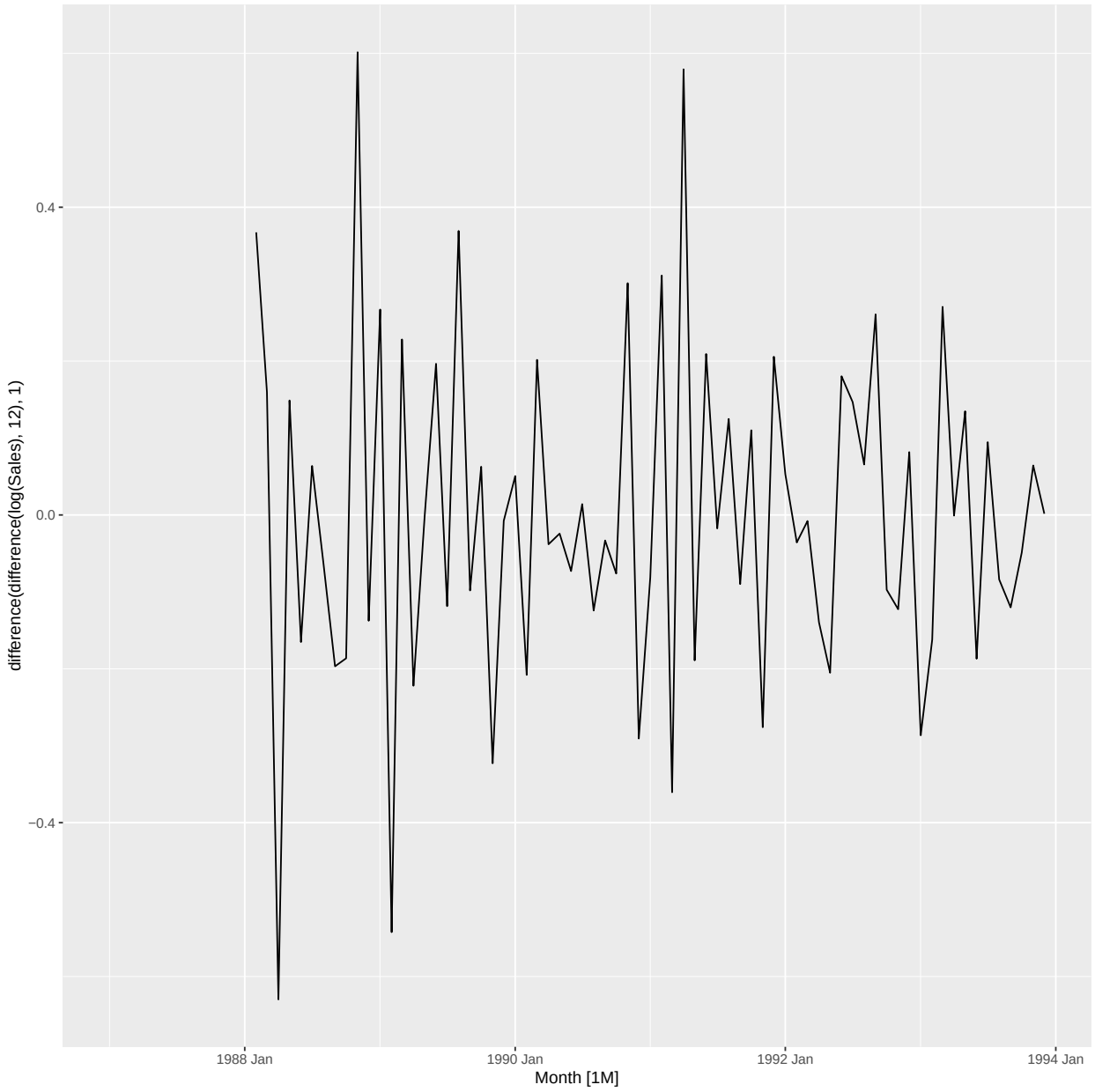
I am curious to see how well this would do with a difference of 12:

```
## Warning: Removed 12 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



After seeing this chart, I was interested to see it differenced once more, as it was done during the example and I found that it looks much better. I'm curious as to why the KPSS test didn't recommend this difference. Is this because this is a derivative difference?

```
## Warning: Removed 13 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```



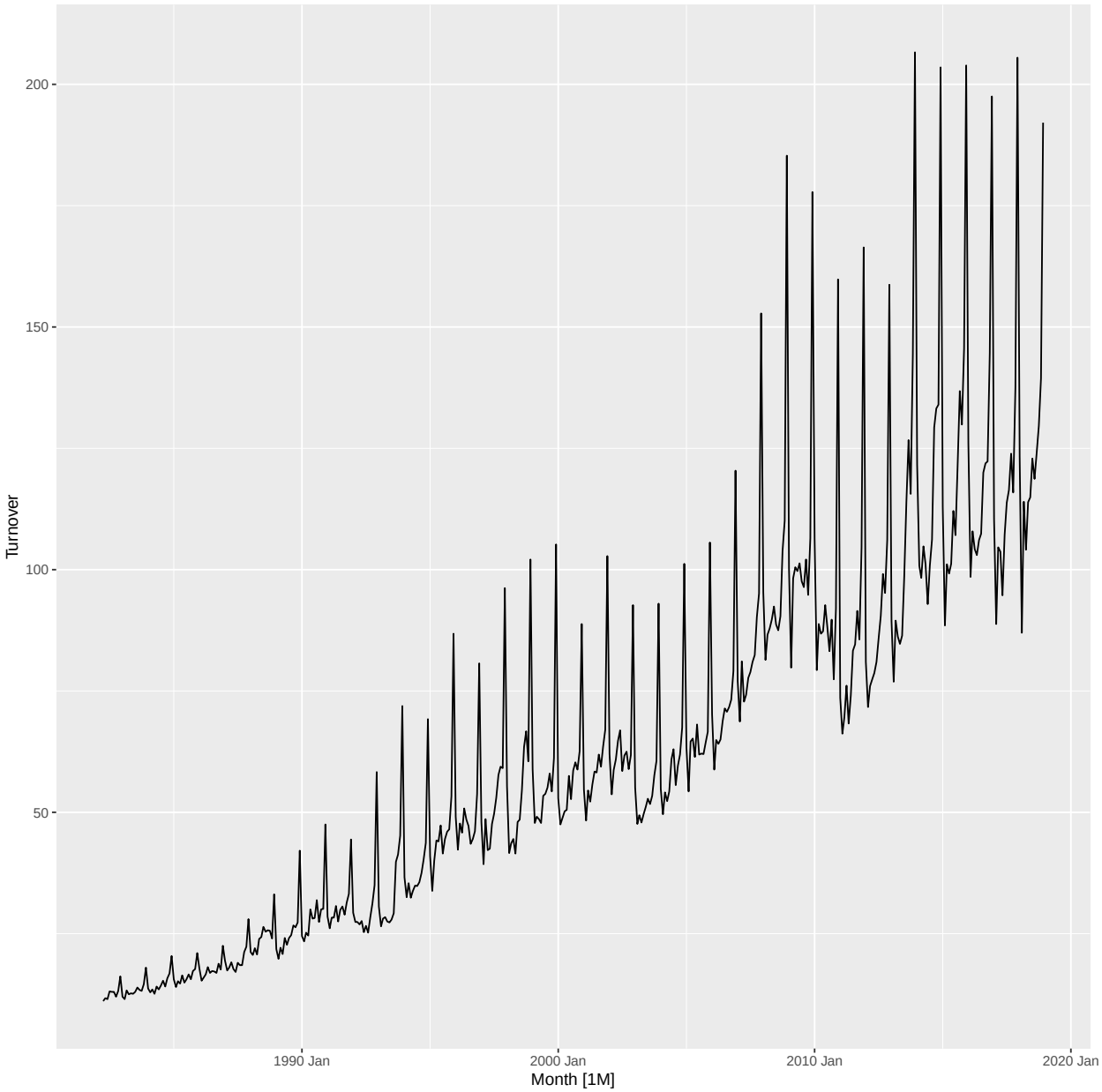
Question 9.5

For your retail data (from Exercise 7 in Section 2.10), find the appropriate order of differencing (after transformation if necessary) to obtain stationary data.

```
## # A tsibble: 6 x 2 [1M]
##   Turnover    Month
##      <dbl>    <mth>
## 1     11.1 1982 Apr
## 2     11.7 1982 May
## 3     11.5 1982 Jun
## 4     13.1 1982 Jul
## 5     13   1982 Aug
```

```
## 6     13   1982 Sep
```

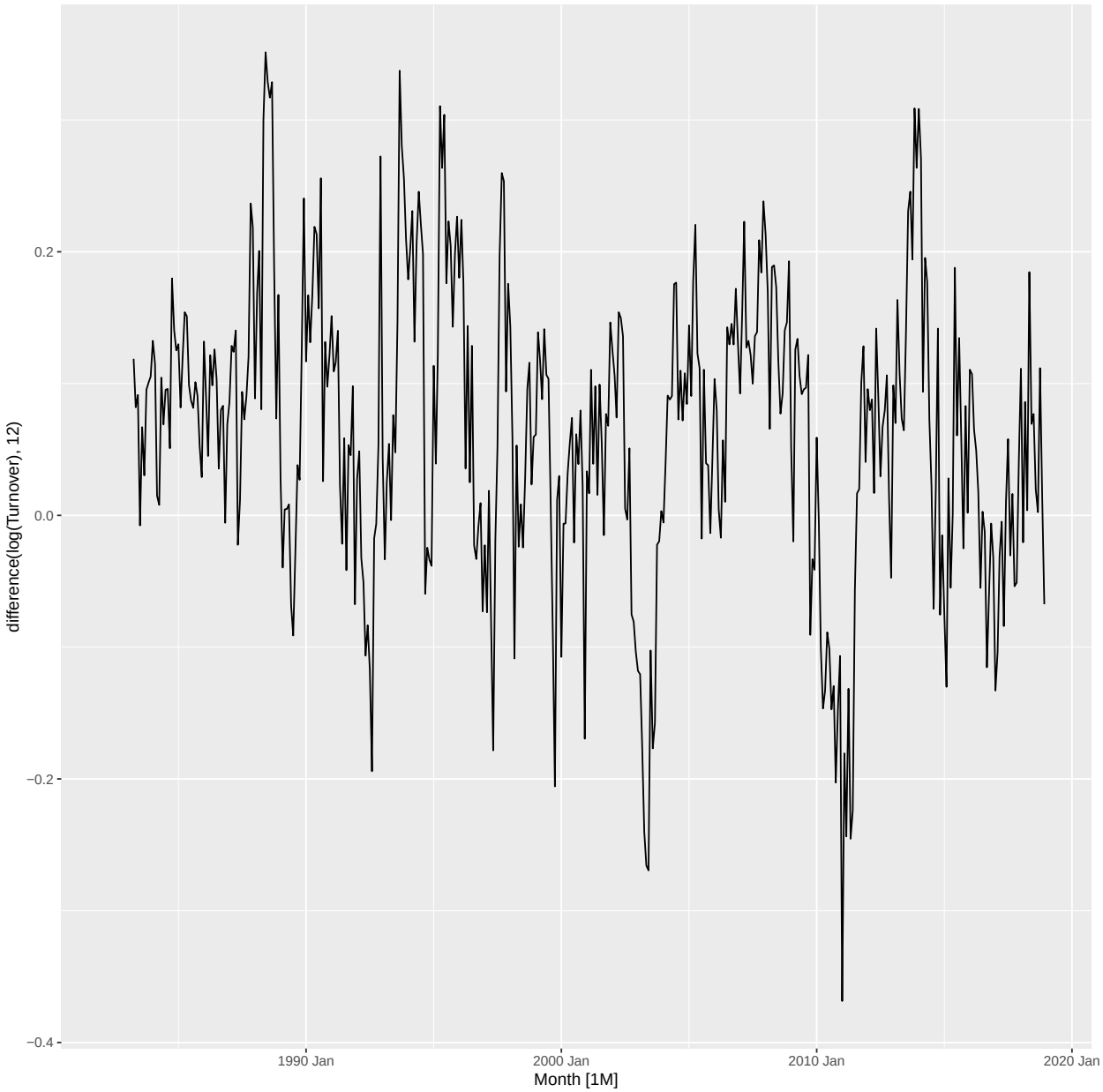
```
## Plot variable not specified, automatically selected `.vars = Turnover`
```



Firstly, let's see if we'll see if we need to perform a box-cox transform.

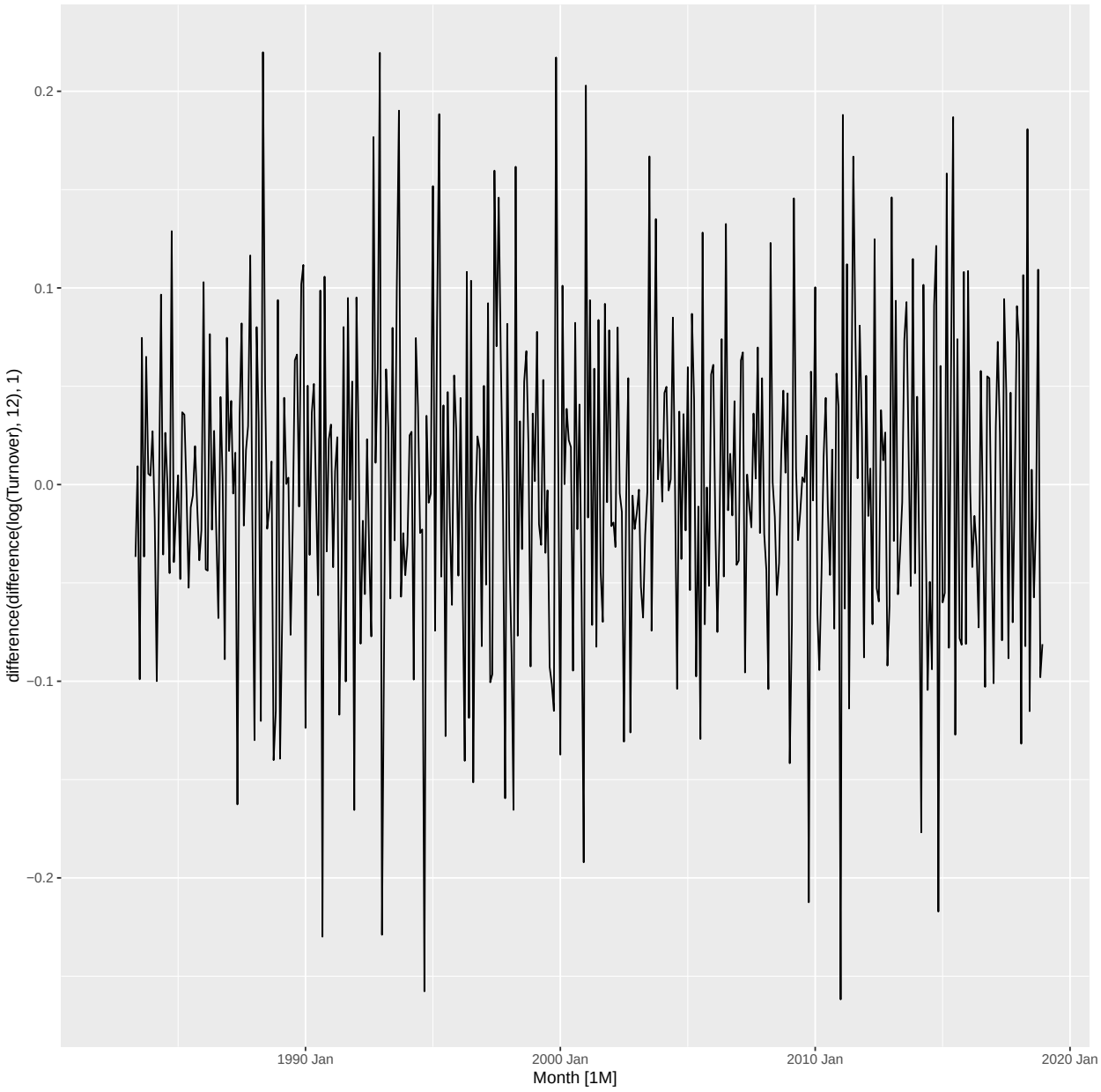
A λ of -0.2 suggests that we should perform the $log(y)$ transformation. With that, we'll perform a difference that is centered around the seasonal duration (12):

```
## Warning: Removed 12 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



Again, although this data seems pretty stationary it seems to have little trends within it. It is possible to perform another difference on the differenced data.

```
## Warning: Removed 13 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```



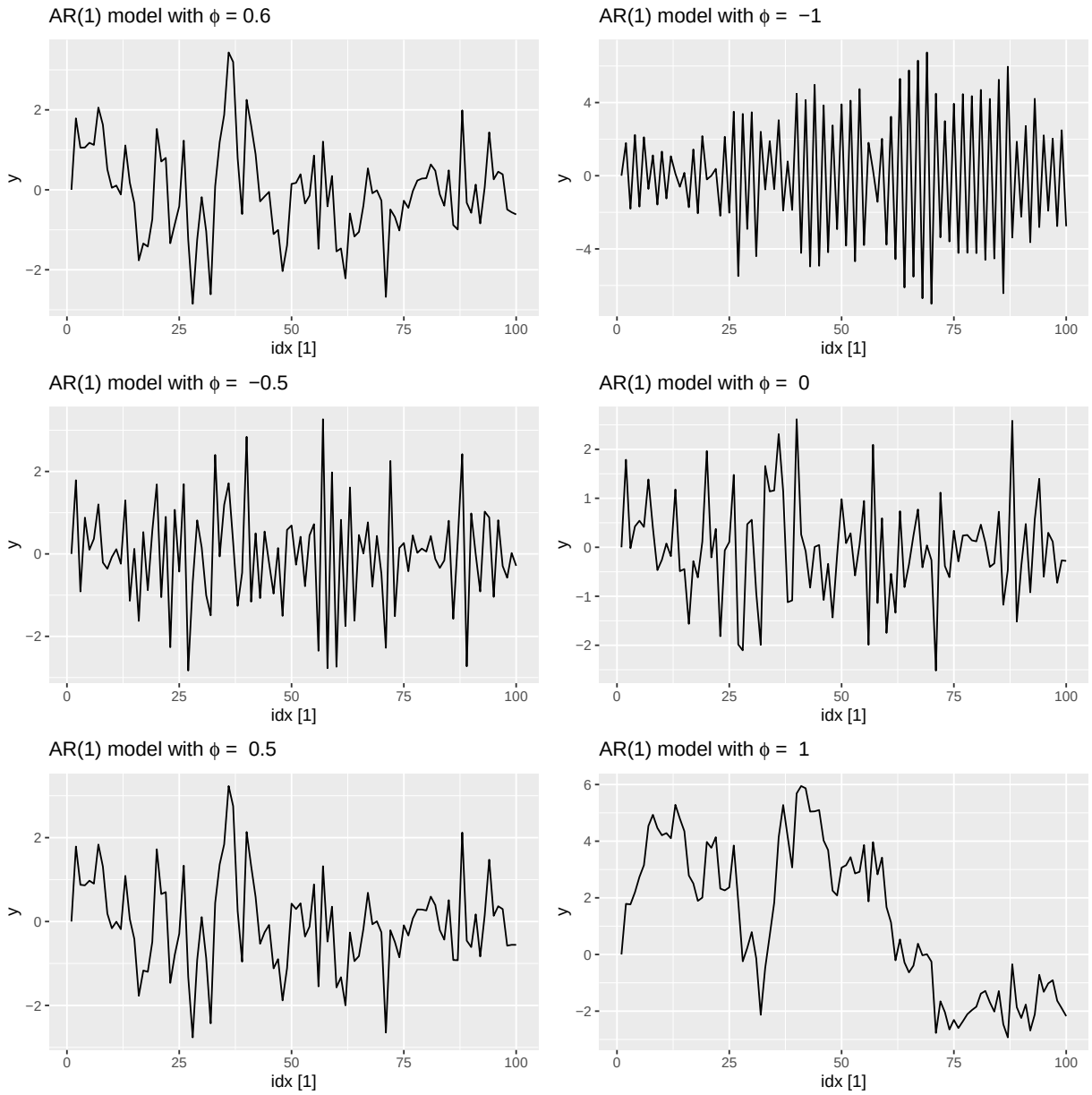
Question 9.6

Simulate and plot some data from simple ARIMA models.

1. Use the following R code to generate data from an AR(1) model with $\phi = 0.6$ and $\sigma^2 = 1$.The process starts with $y_1 = 0$.
2. Produce a time plot for the series. How does the plot change as you change ϕ_1?

```
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y`
```

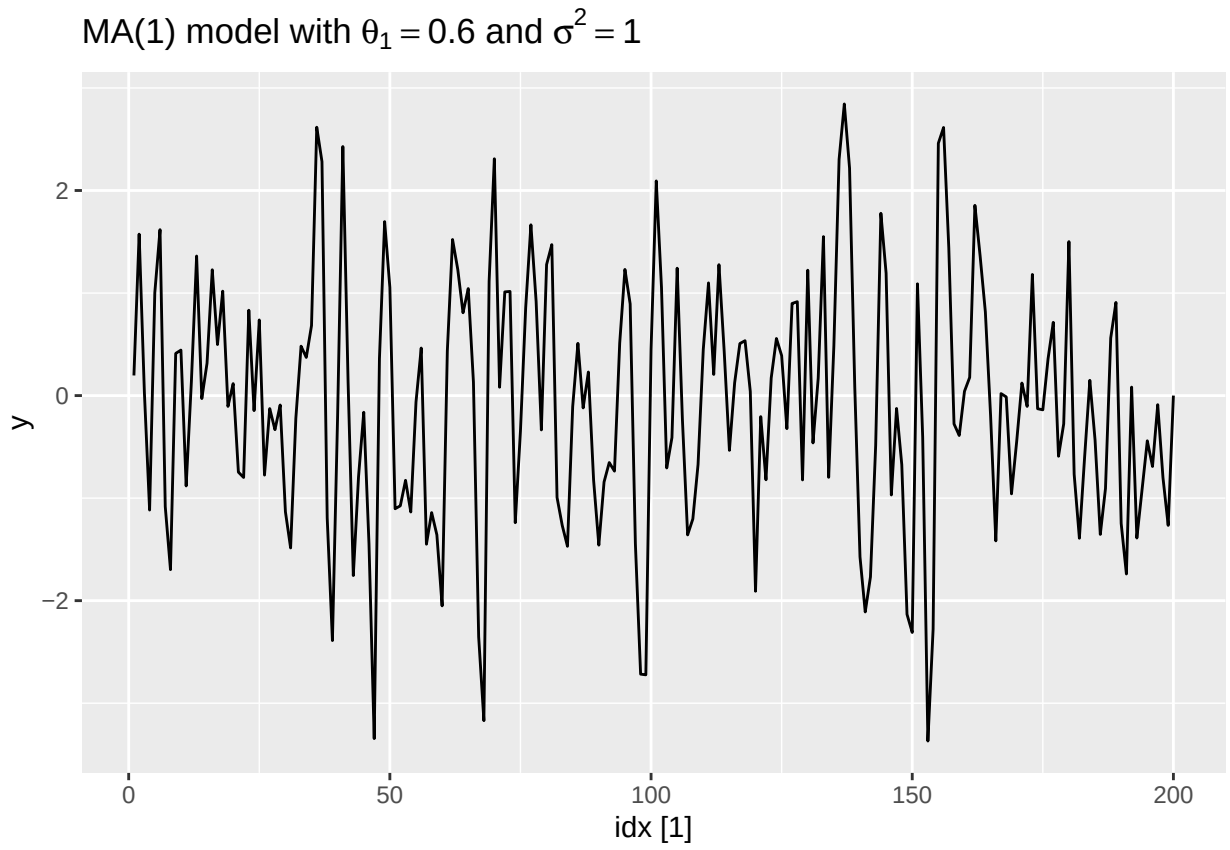
```
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y`
```



Each of these phi values produce a very different plot. When $\phi = -1$ the plot seems to oscillate around 0 and when ϕ is in -0.6, -0.5, 0, and 0.5 it looks like a noise. When $\phi = 1$ it resembles a timeseries with a trend.

3. Write your own code to generate data from an MA(1) model with $\theta_1 = 0.6$ and $\sigma^2 = 1$.

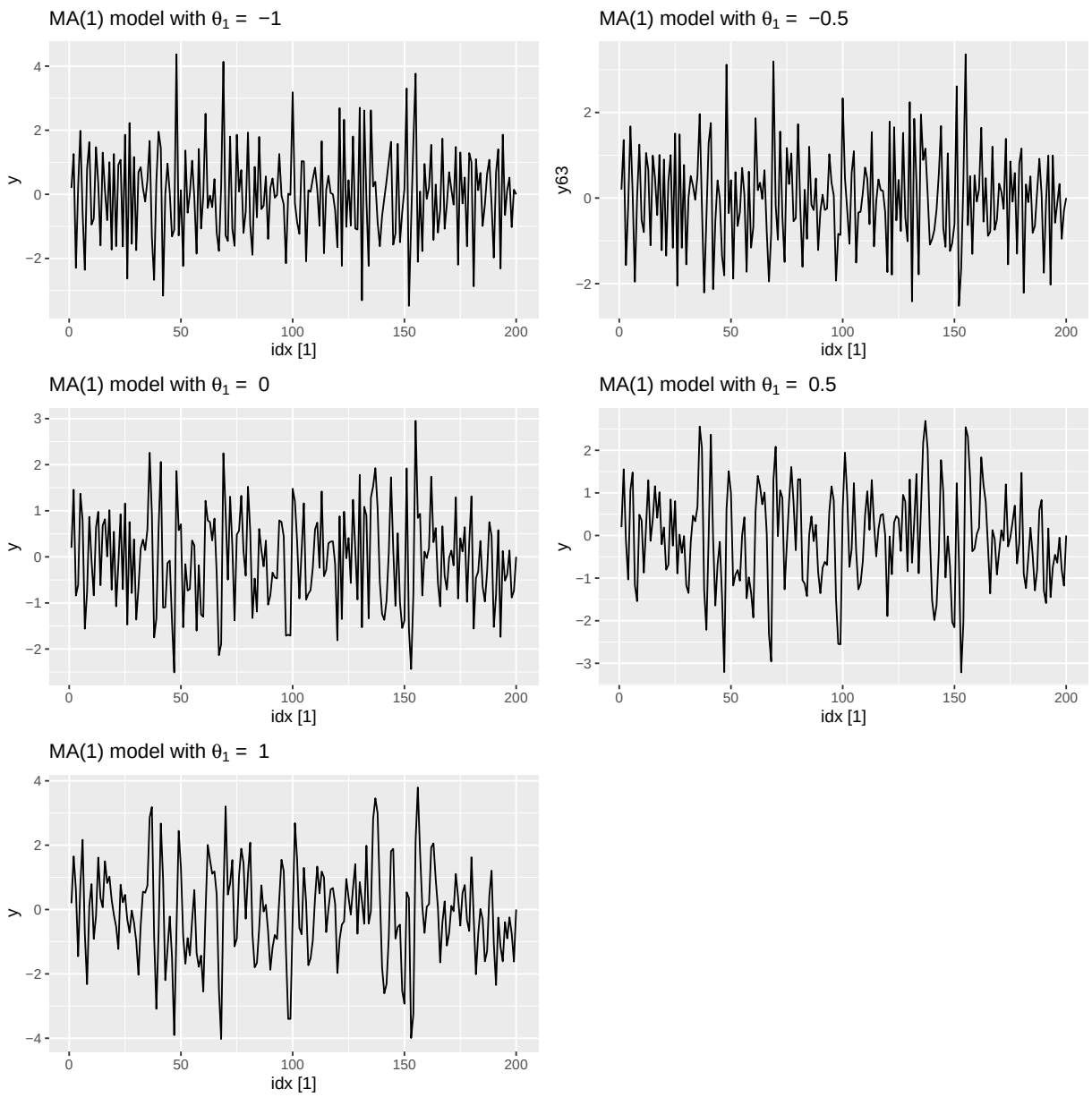
```
## Plot variable not specified, automatically selected `.vars = y`
```



4. Produce a time plot for the series. How does the plot change as you change θ_1?

   The MA(1) model must be between -1 and 1, therefore:

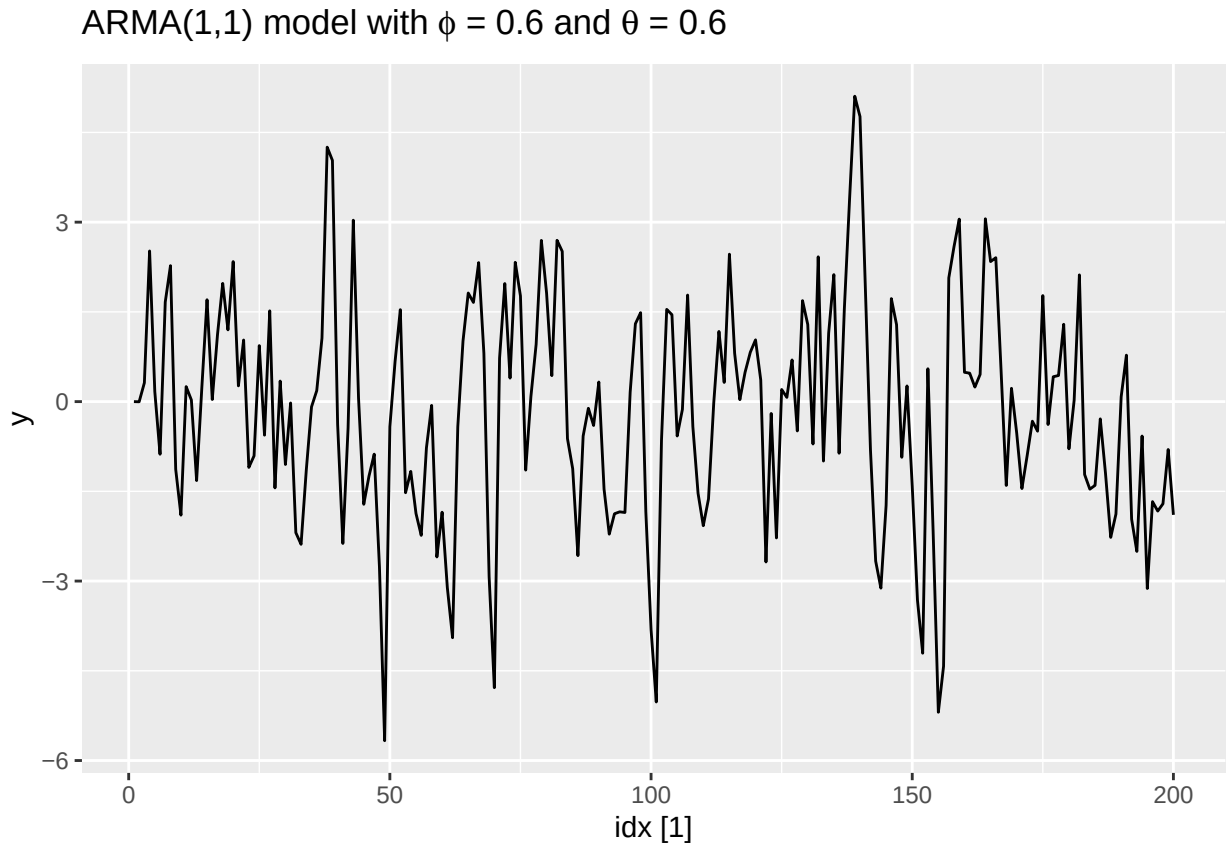
```
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y63`
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y`
## Plot variable not specified, automatically selected `.vars = y`
```



When θ is -1 or -0.5, the data seems to be more erratic and as it increases and approaches 1 the data seems to become less erratic. Aside from that, the data appears to be stationary.

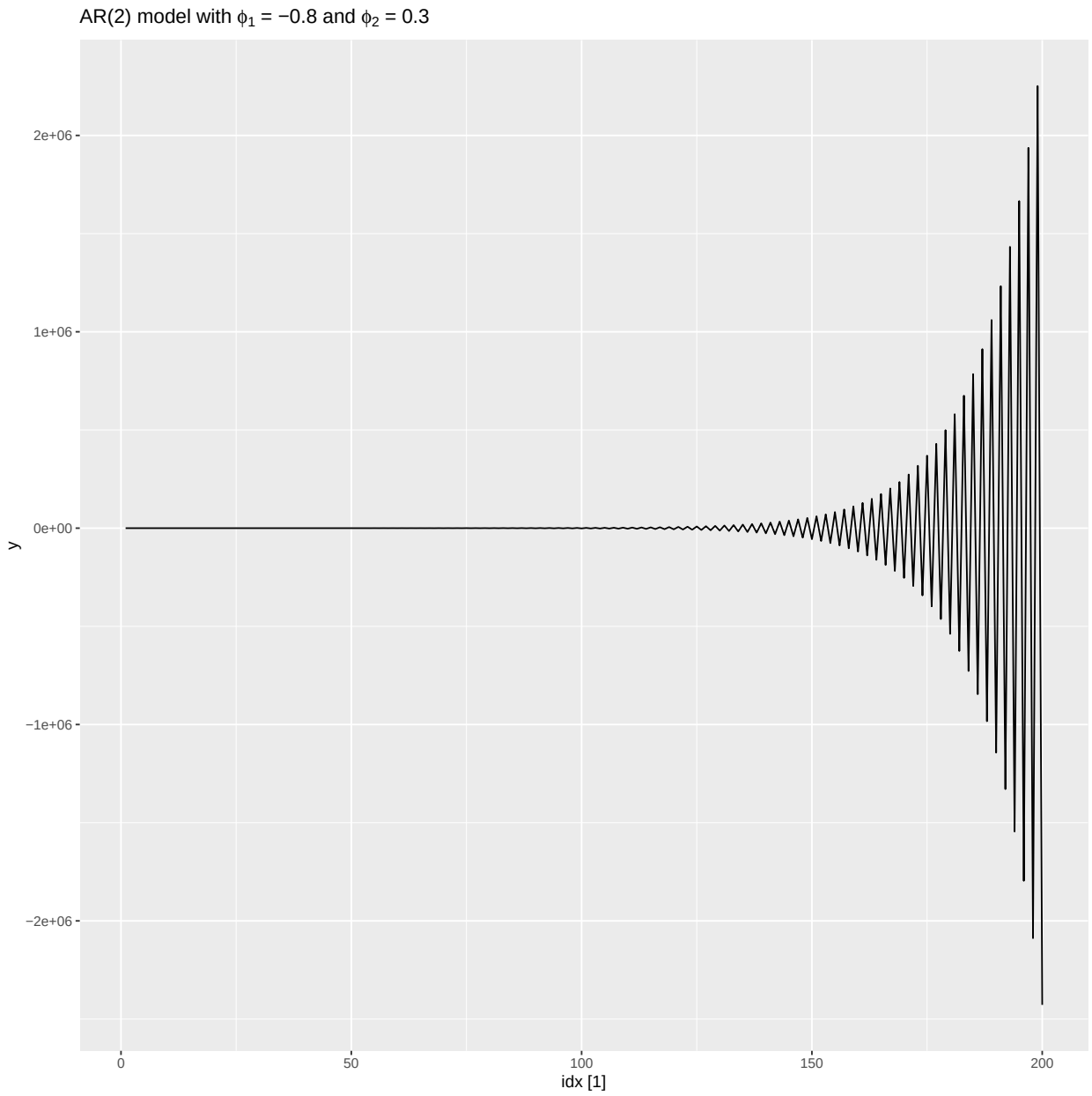
5. Generate data from an ARMA(1,1) model with $\phi_1 = 0.6$, $\theta_1 = 0.6$, and $\sigma^2 = 1$.

```
## Plot variable not specified, automatically selected `.vars = y`
```



5. Generate data from an AR(2) model with $\phi_1 = -0.8$, $\phi_2 = 0.3$, and $\sigma^2 = 1$. (Note that these parameters will give a non-stationary series.)

```
## Plot variable not specified, automatically selected `.vars = y`
```



6. Graph the latter two series and compare them.

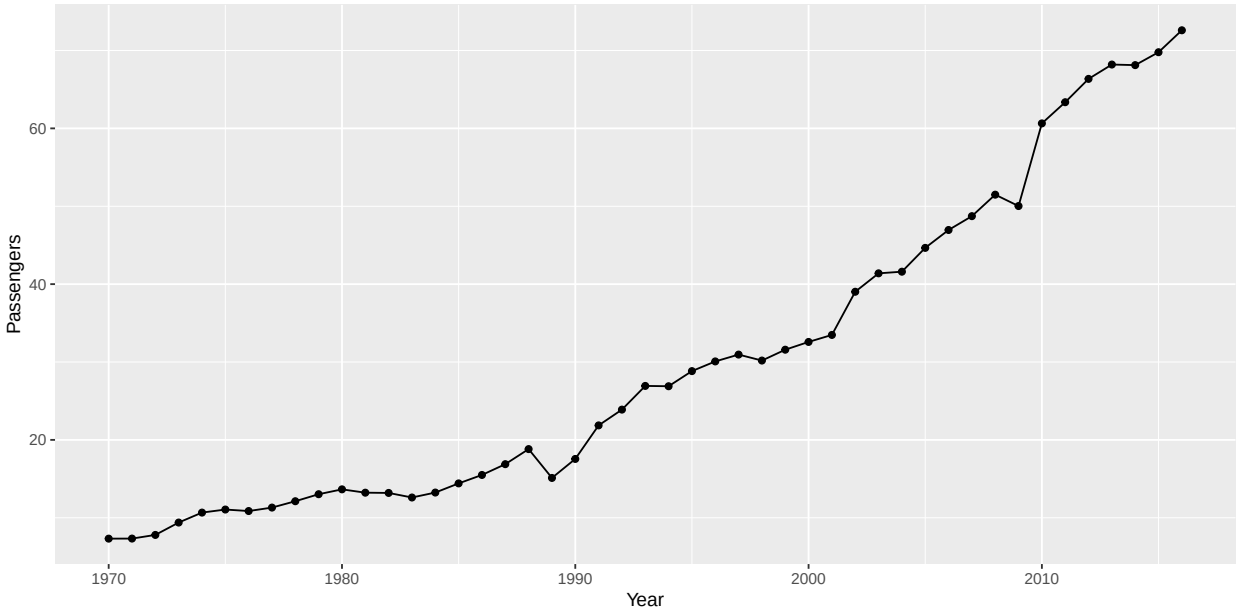
   The `ARMA(1,1)` seems stationary and like noise while the AR(2) model seems like an unstable, growing in magnitude as time goes on.

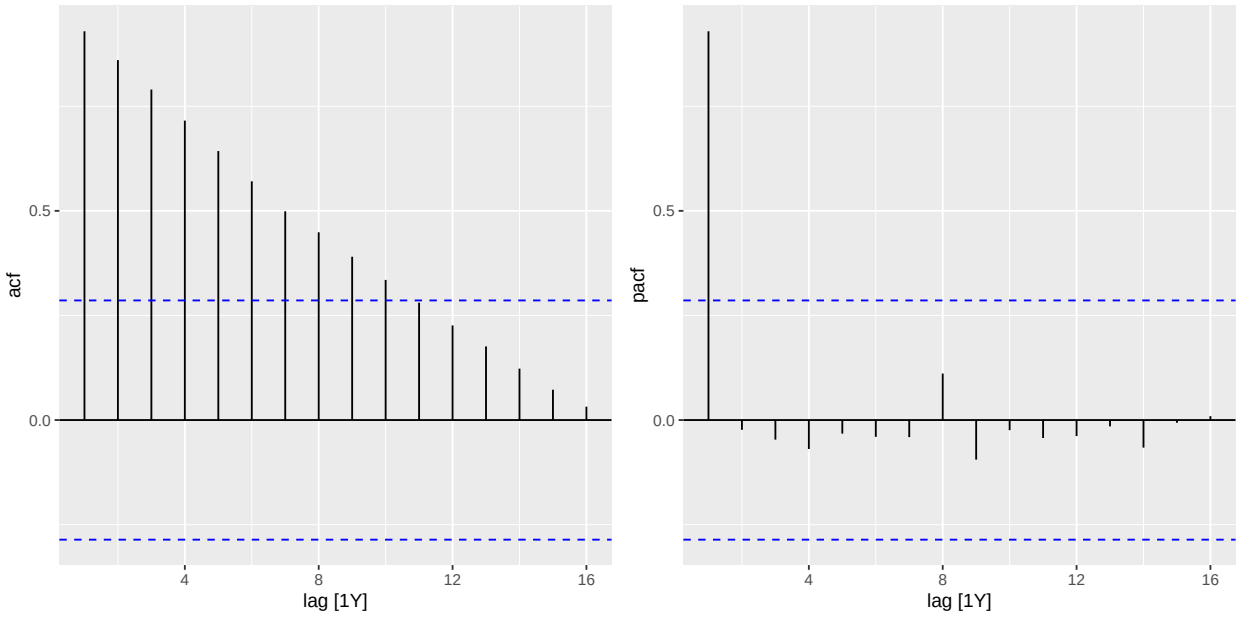
Question 9.7

Consider `aus_airpassengers`, the total number of passengers (in millions) from Australian air carriers for the period 1970-2011.

1. Use `ARIMA()` to find an appropriate ARIMA model. What model was selected. Check that the residuals look like white noise. Plot forecasts for the next 10 periods.

```
## # A tsibble: 6 x 2 [1Y]
##    Year Passengers
##   <dbl>      <dbl>
## 1  1970       7.32
## 2  1971       7.33
## 3  1972       7.80
## 4  1973       9.38
## 5  1974      10.7
## 6  1975      11.1
```

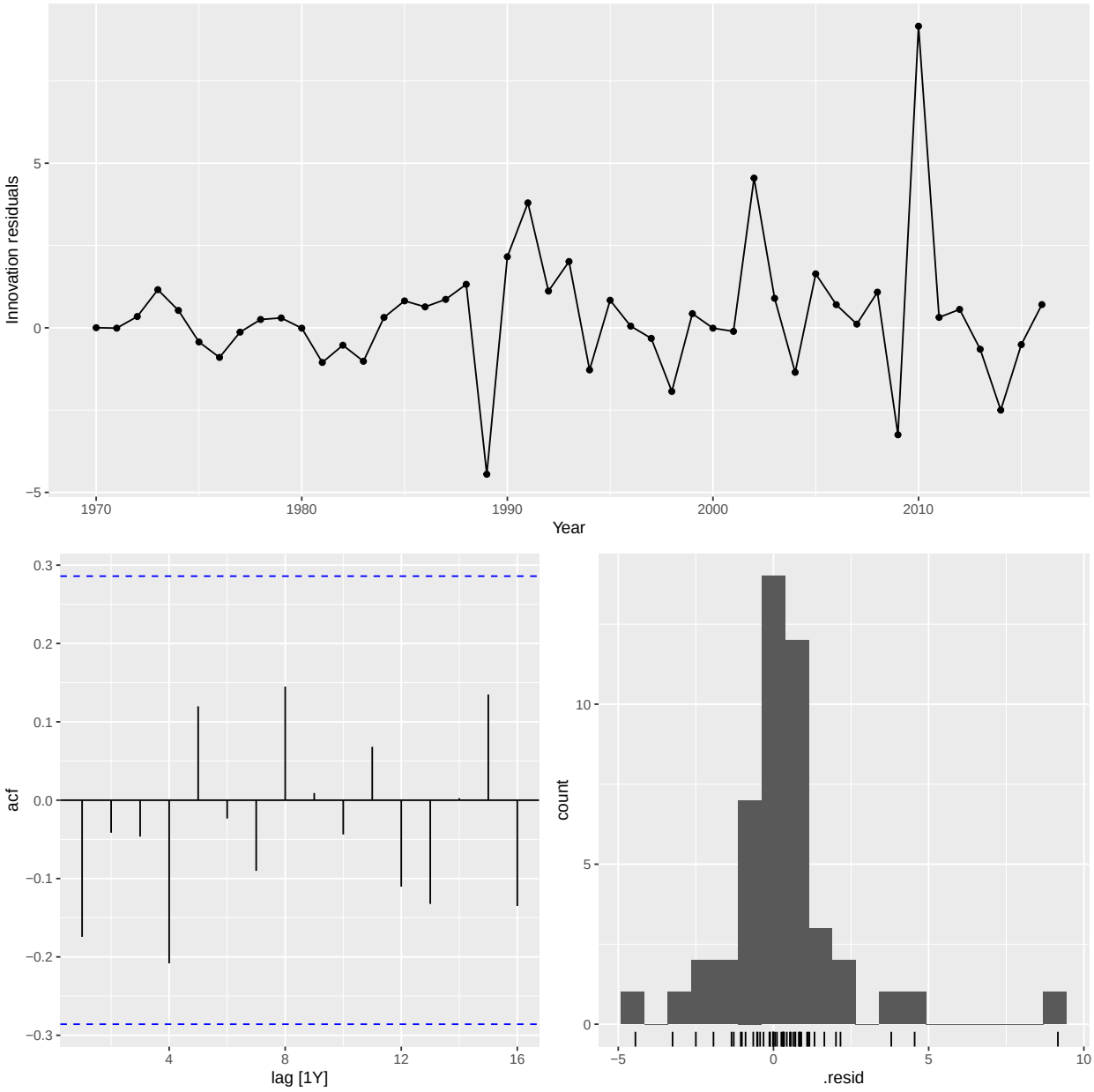





```
## Series: Passengers
## Model: ARIMA(0,2,1)
##
## Coefficients:
```

```
##           ma1
##       -0.8963
## s.e.   0.0594
##
## sigma^2 estimated as 4.308:  log likelihood=-97.02
## AIC=198.04   AICc=198.32   BIC=201.65
```

An ARIMA(0, 2, 1) was selected.

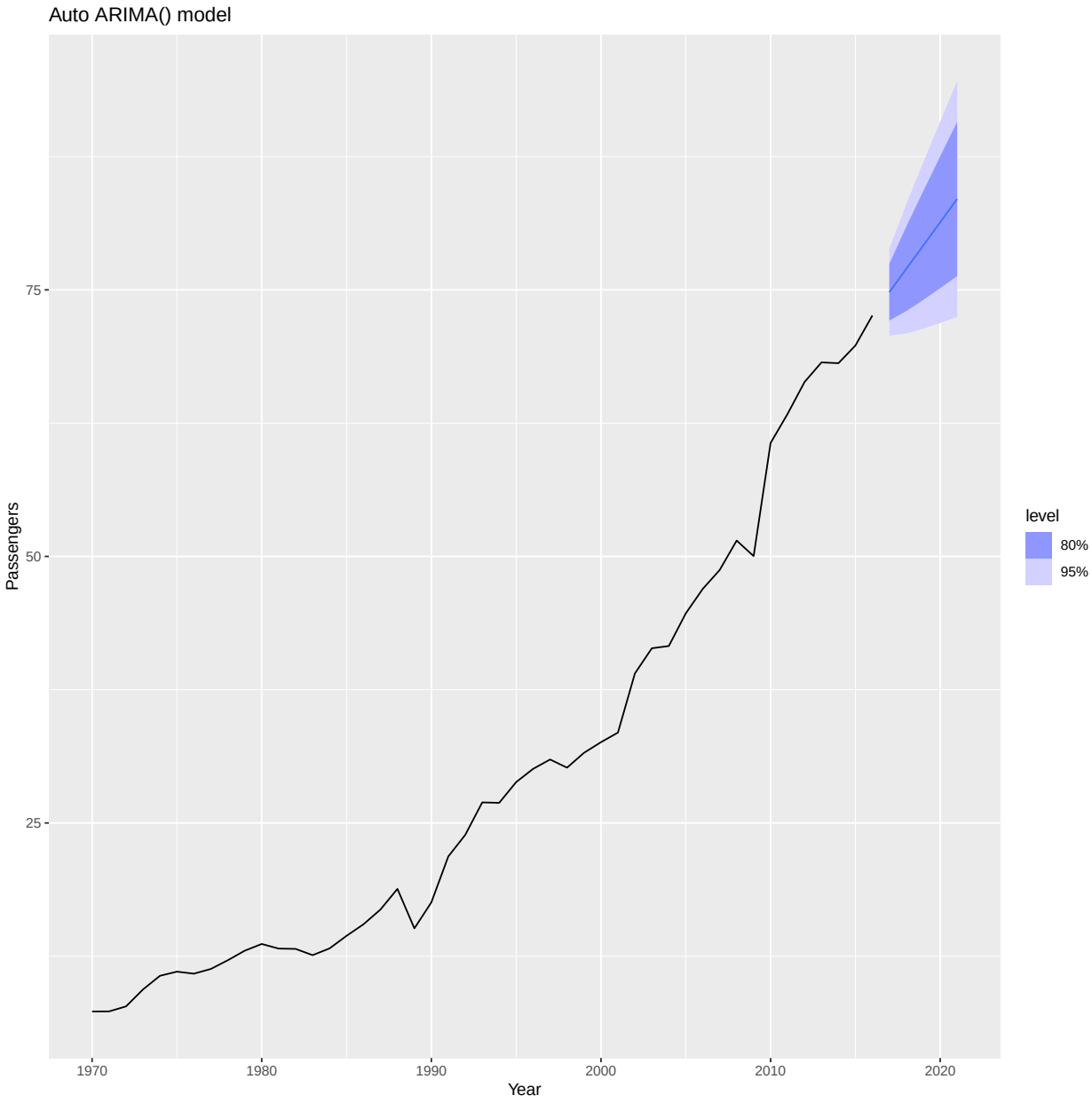


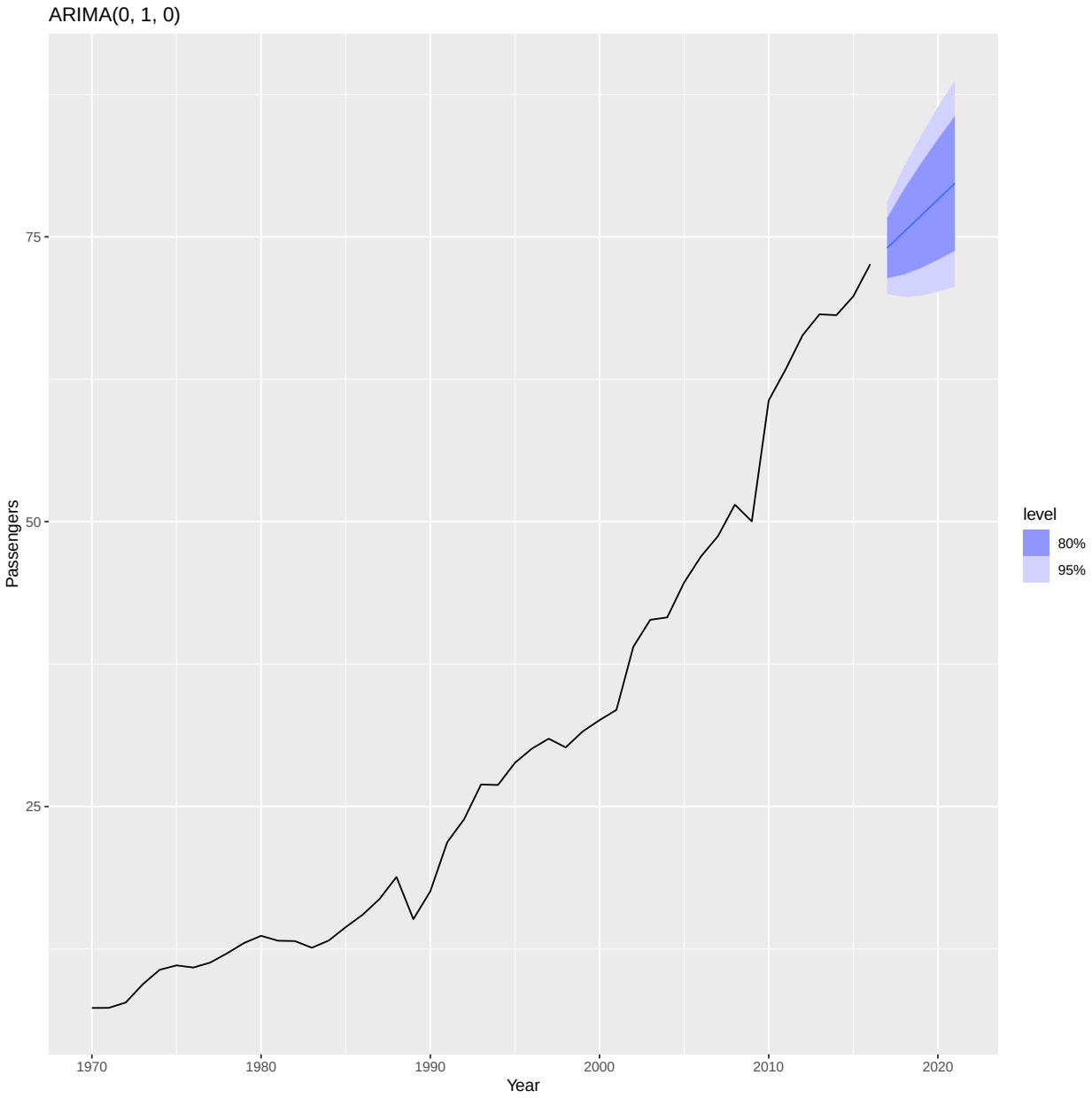
The above charts show the residuals where we can see on the histogram that they seem fairly normally distributed, the acf has values all below the critical value and the line plot seems to be white noise around 0, i.e. is stationary.

2. Write the model in terms of the backshift operator.

$$(1-B)^2 y_t = (1+(-0.896)B)\epsilon_t$$

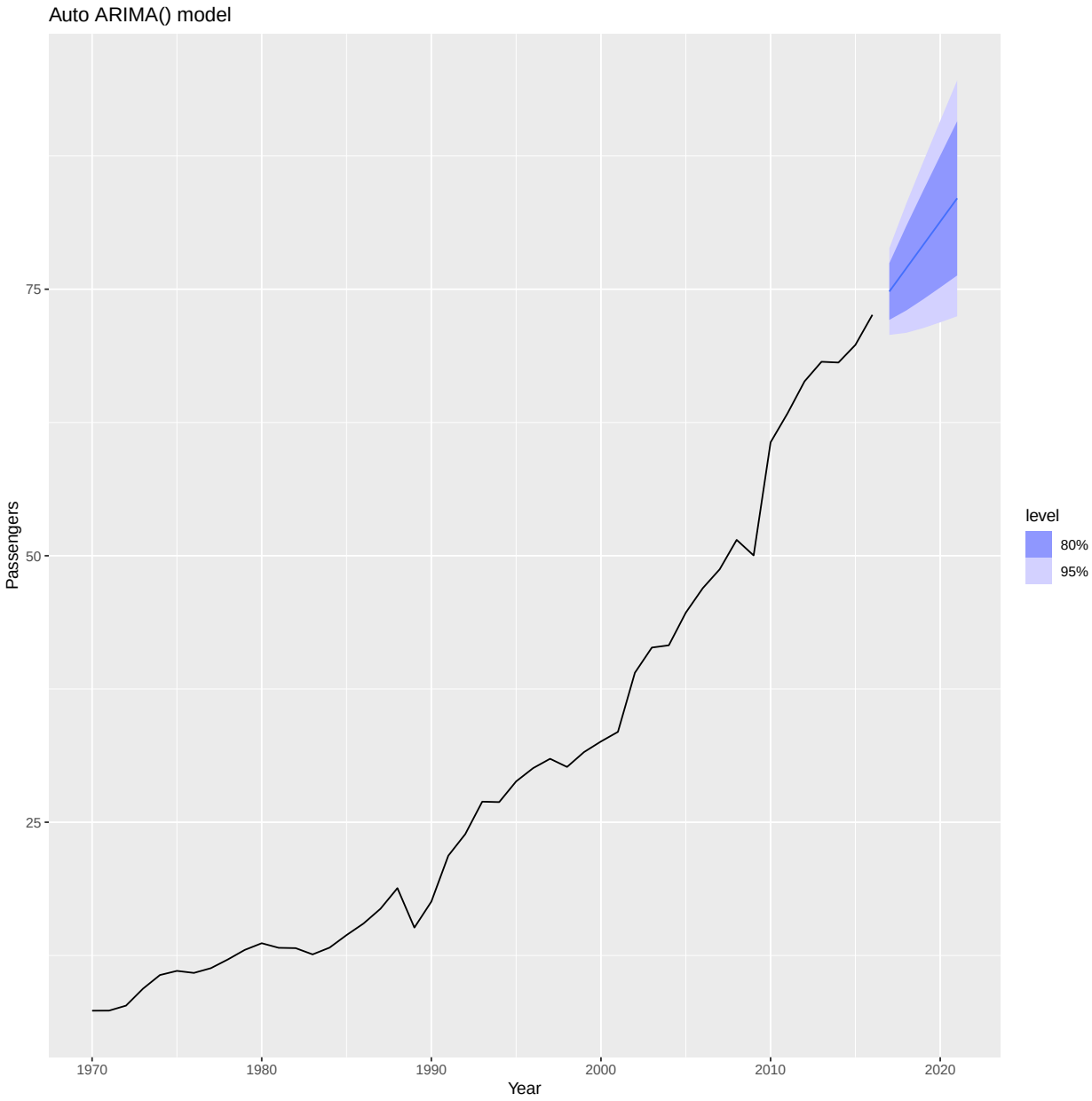
3. Plot forecasts from an ARIMA(0,1,0) model with drift and compare these to part a.





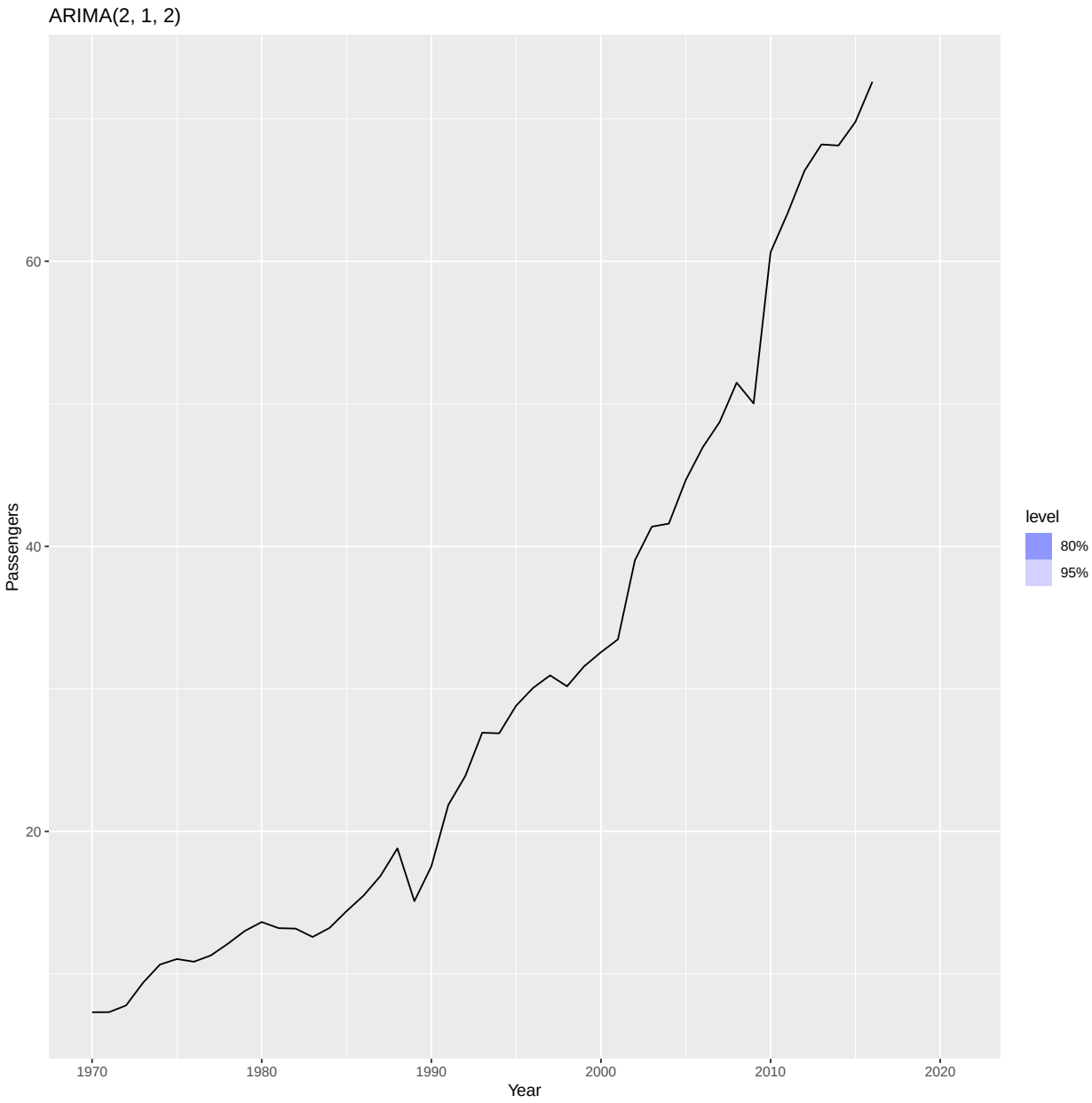
The ARIMA(0, 1, 0) model seems to have a lower slope although much of the prediction interval seems to overlap between the two models.

4. Plot forecasts from an ARIMA(2,1,2) model with drift and compare these to parts a and c. Remove the constant and see what happens.

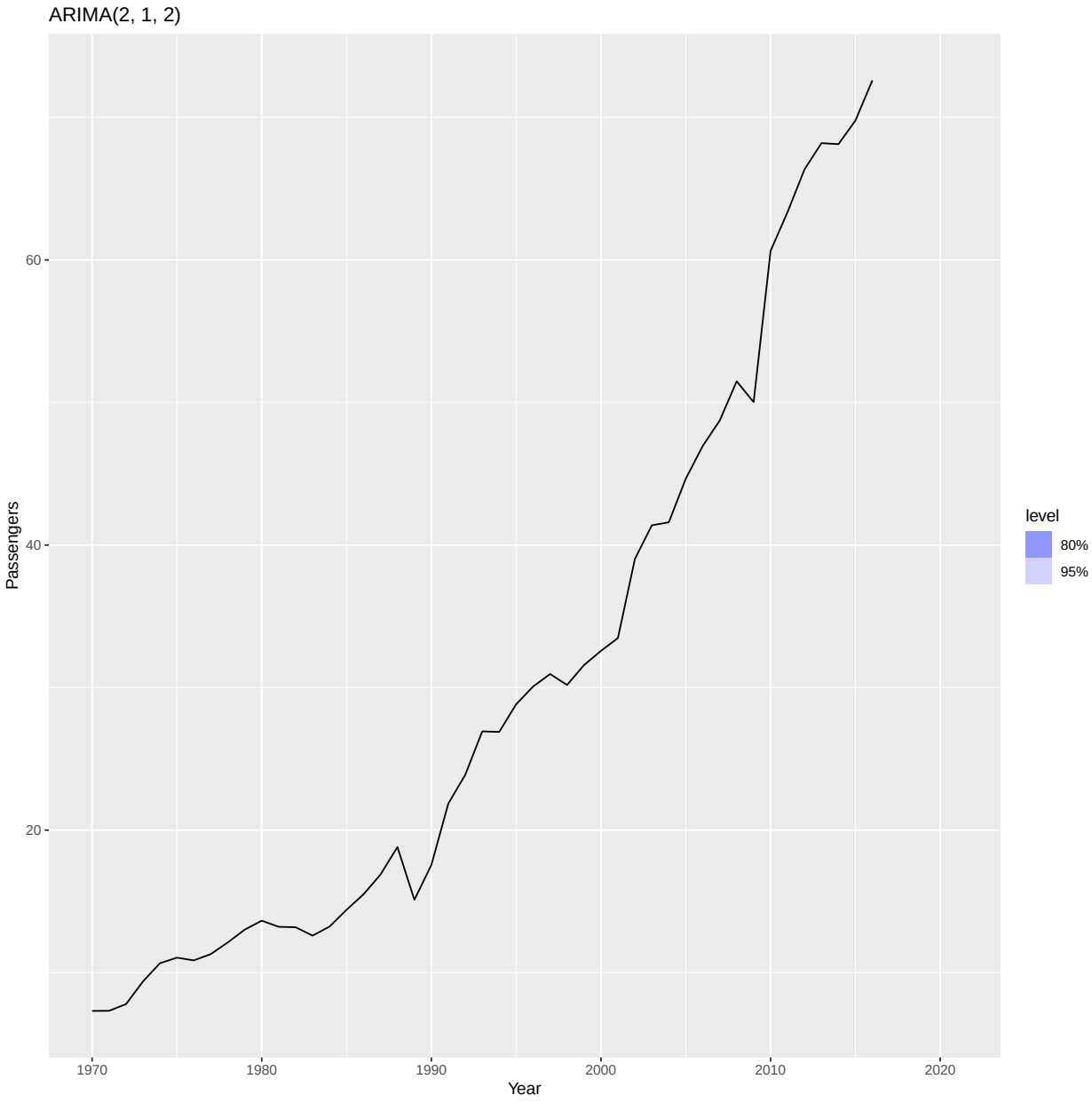



```
## Warning: It looks like you're trying to fully specify your ARIMA model but have not said if a constar
## You can include a constant using `ARIMA(y~1)` to the formula or exclude it by adding `ARIMA(y~0)`.

## Warning: 1 error encountered for arima_2_1_2
## [1] Could not find an appropriate ARIMA model.
## This is likely because automatic selection does not select models with characteristic roots that may
## For more details, refer to https://otexts.com/fpp3/arima-r.html#plotting-the-characteristic-roots

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_line()`).
```

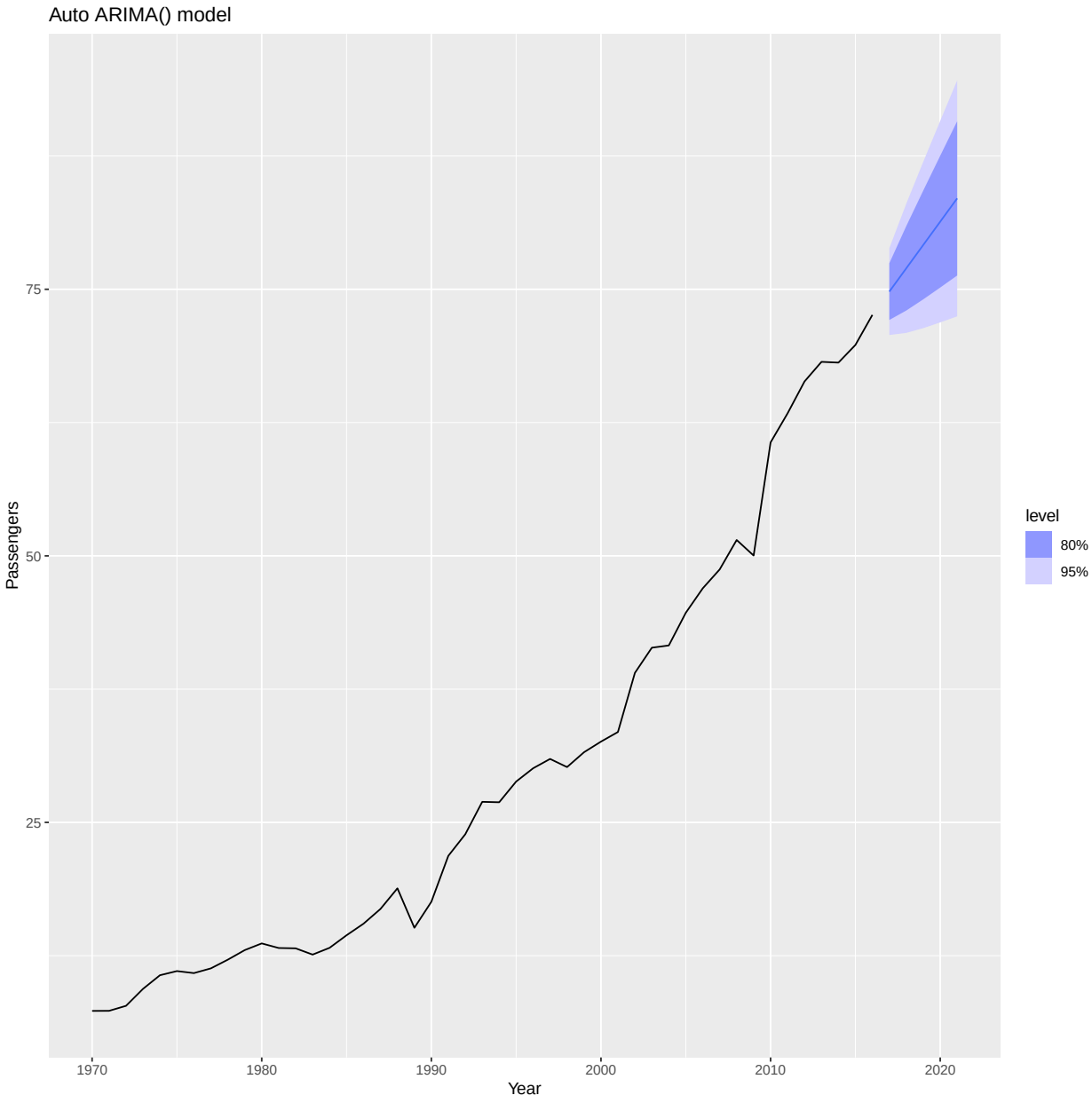


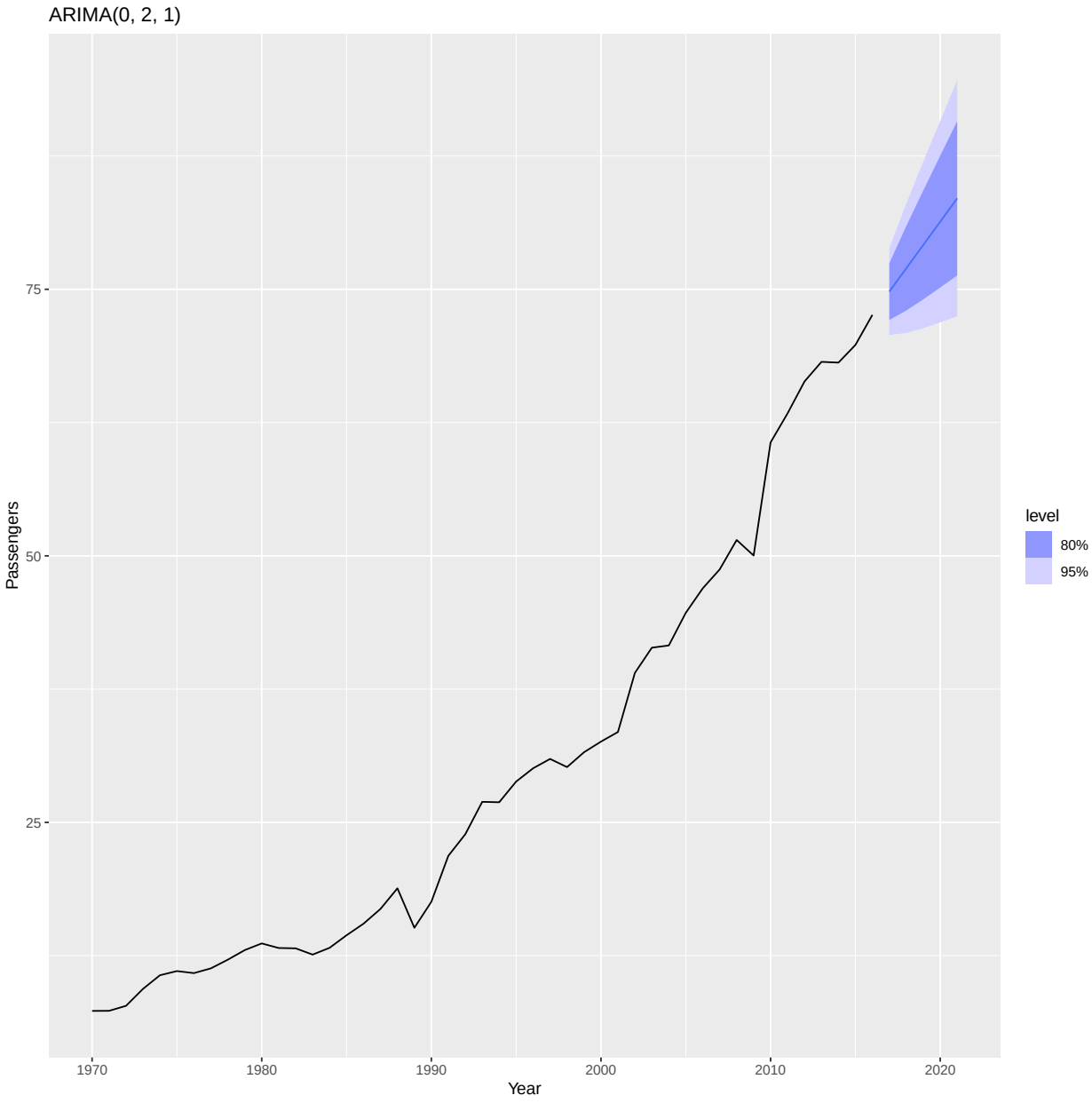

```
## Warning: 1 error encountered for arima_2_1_2
## [1] non-stationary AR part from CSS

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf

## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



The `ARIMA(2,1,2)` model given an error as the AR part was non stationary.

5. Plot forecasts from an ARIMA(0,2,1) model with a constant. What happens?


Auto ARIMA() model
Passengers
75
50
25
1970
1980
1990
2000
2010
2020
Year
level
80%
95%



This is the same model that was automatically selected by calling `ARIMA()` so the results are the same.

Question 9.8

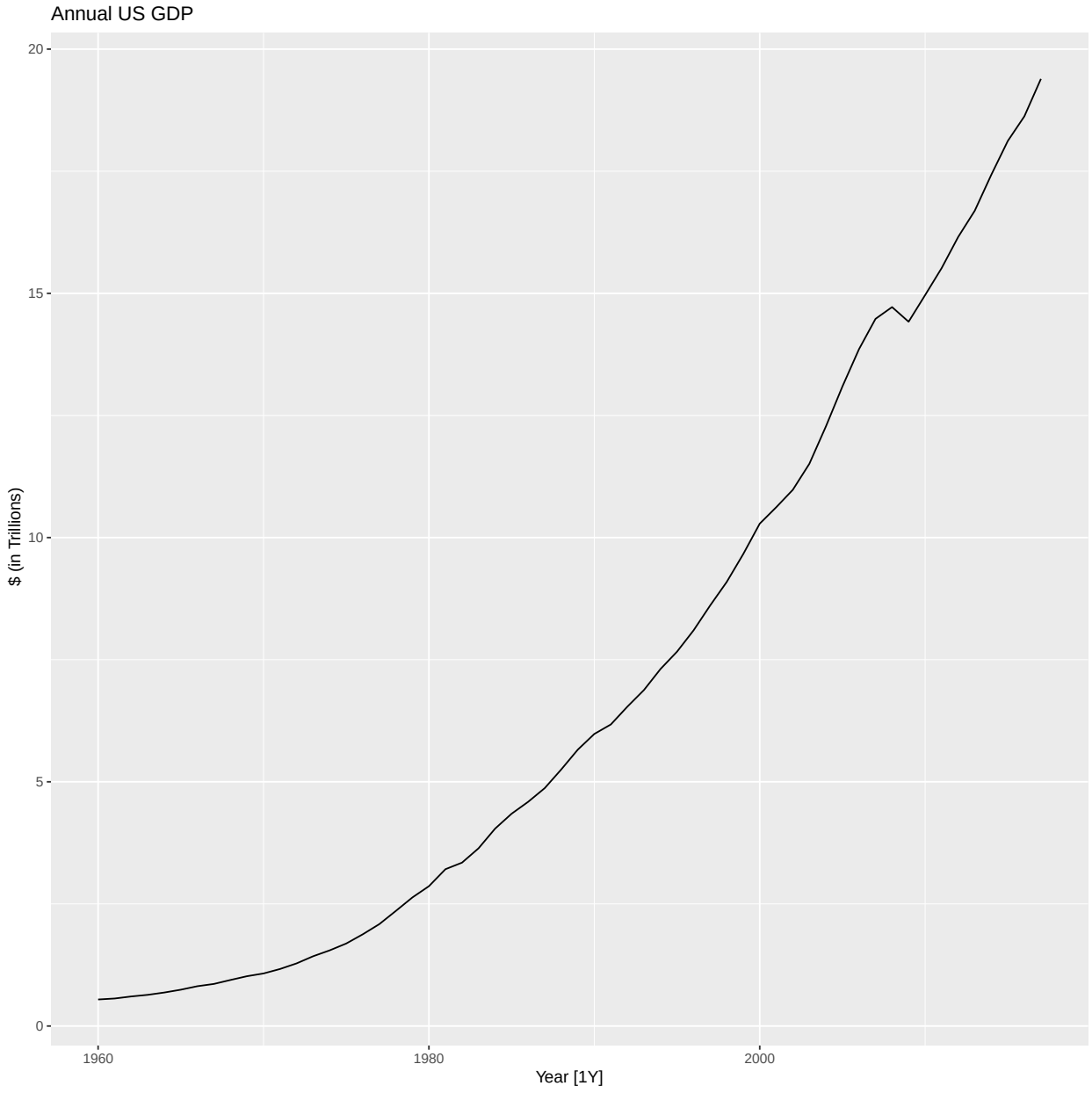
For the United States GDP series (from `global_economy`):

1. if necessary, find a suitable Box-Cox transformation for the data;

```
## # A tsibble: 6 x 9 [1Y]
## # Key:       Country [1]
##   Country     Code   Year         GDP Growth   CPI Imports Exports Population
```

```
##   <fct>       <fct> <dbl>       <dbl>  <dbl> <dbl>   <dbl>   <dbl>      <dbl>
## 1 Afghanistan AFG    1960  537777811.     NA    NA    7.02    4.13    8996351
## 2 Afghanistan AFG    1961  548888896.     NA    NA    8.10    4.45    9166764
## 3 Afghanistan AFG    1962  546666678.     NA    NA    9.35    4.88    9345868
## 4 Afghanistan AFG    1963  751111191.     NA    NA   16.9     9.17    9533954
## 5 Afghanistan AFG    1964  800000044.     NA    NA   18.1     8.89    9731361
## 6 Afghanistan AFG    1965 1006666638.     NA    NA   21.4    11.3     9938414
```

```
## Plot variable not specified, automatically selected `.vars = GDP`
```



Because there is no seasonality to the data, I believe that a Box-Cox transform is unnecessary.

2. fit a suitable ARIMA model to the transformed data using `ARIMA()`;

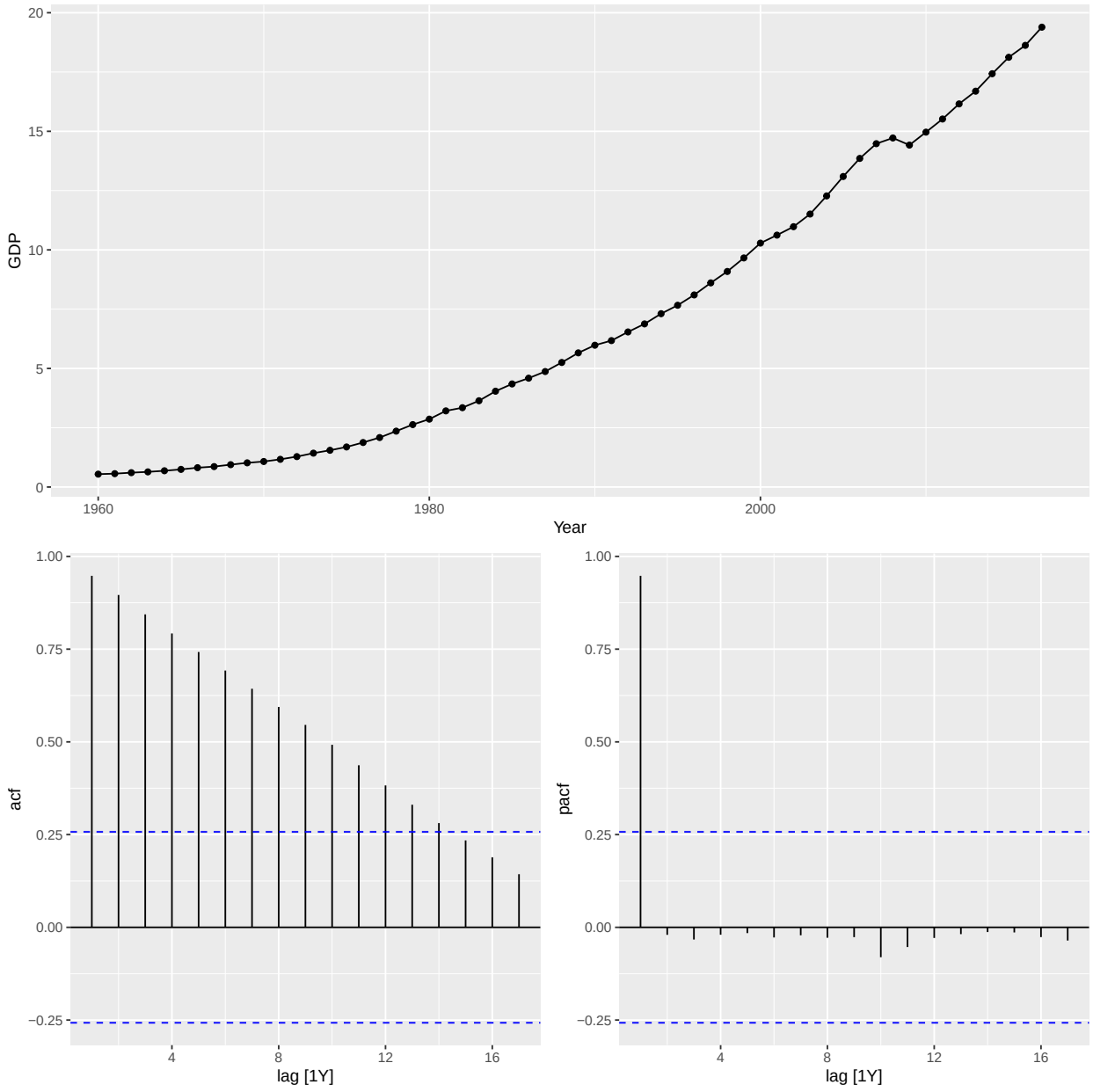
I'll do so by seeing what the `ARIMA()` function recommends:

```
## Series: GDP
## Model: ARIMA(0,2,2)
##
## Coefficients:
##           ma1      ma2
##       -0.4206  -0.3048
## s.e.   0.1197   0.1078
##
## sigma^2 estimated as 0.02615:  log likelihood=23.26
## AIC=-40.52   AICc=-40.06   BIC=-34.45
```

The function recommended an ARIMA(0, 2, 2) model.

3. try some other plausible models by experimenting with the orders chosen;

```
## Plot variable not specified, automatically selected `y = GDP`
```

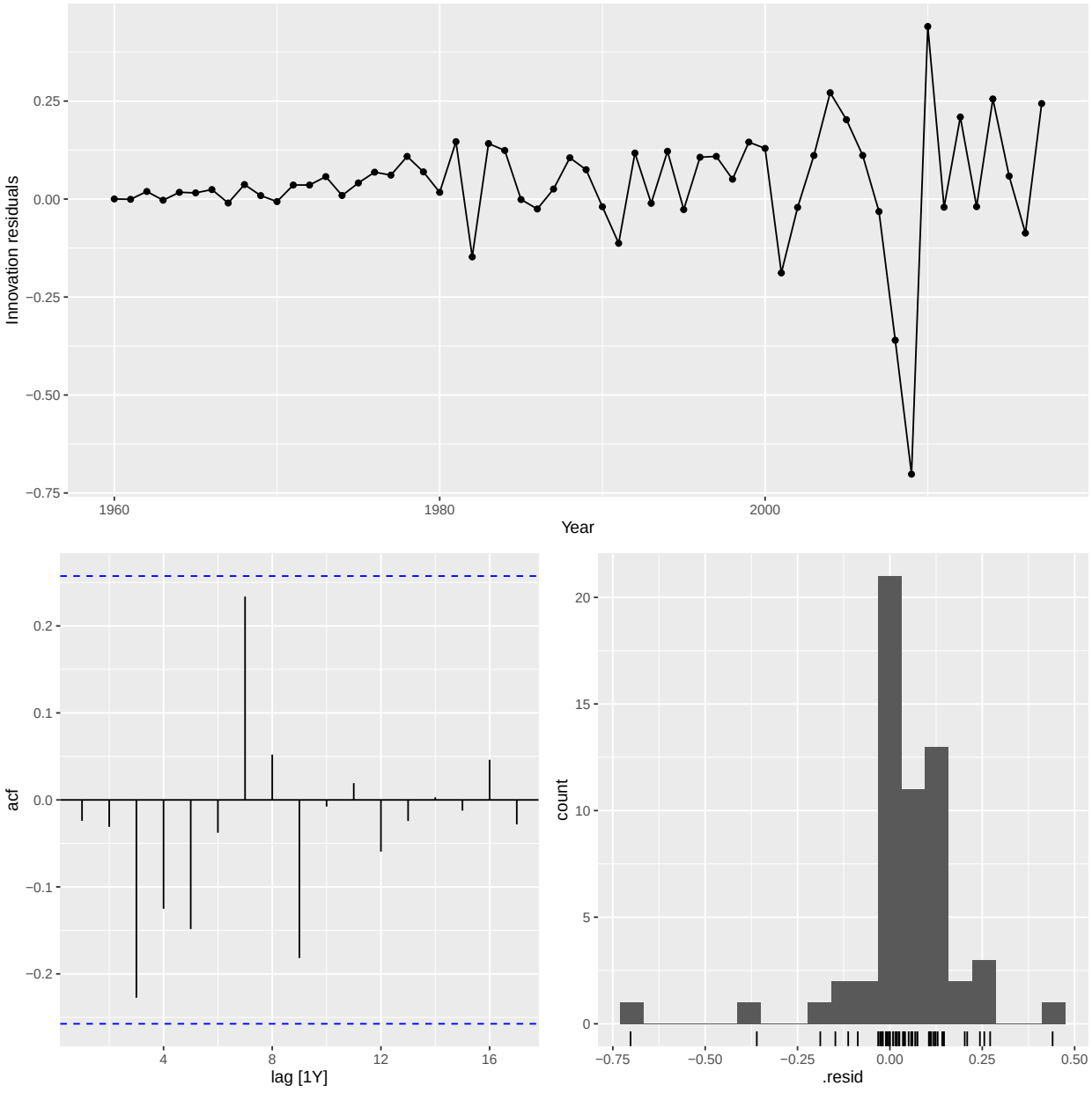


From the above graph, we can see that the ACF seems to decrease linearly and the PACF has an always negative but oscillating component after the first lag.

```
## Series: GDP
## Model: ARIMA(1,2,1)
##
## Coefficients:
##          ar1      ma1
##       0.4105  -0.8391
## s.e.  0.1522   0.0742
##
## sigma^2 estimated as 0.02634:  log likelihood=23.1
## AIC=-40.21   AICc=-39.74   BIC=-34.13
```

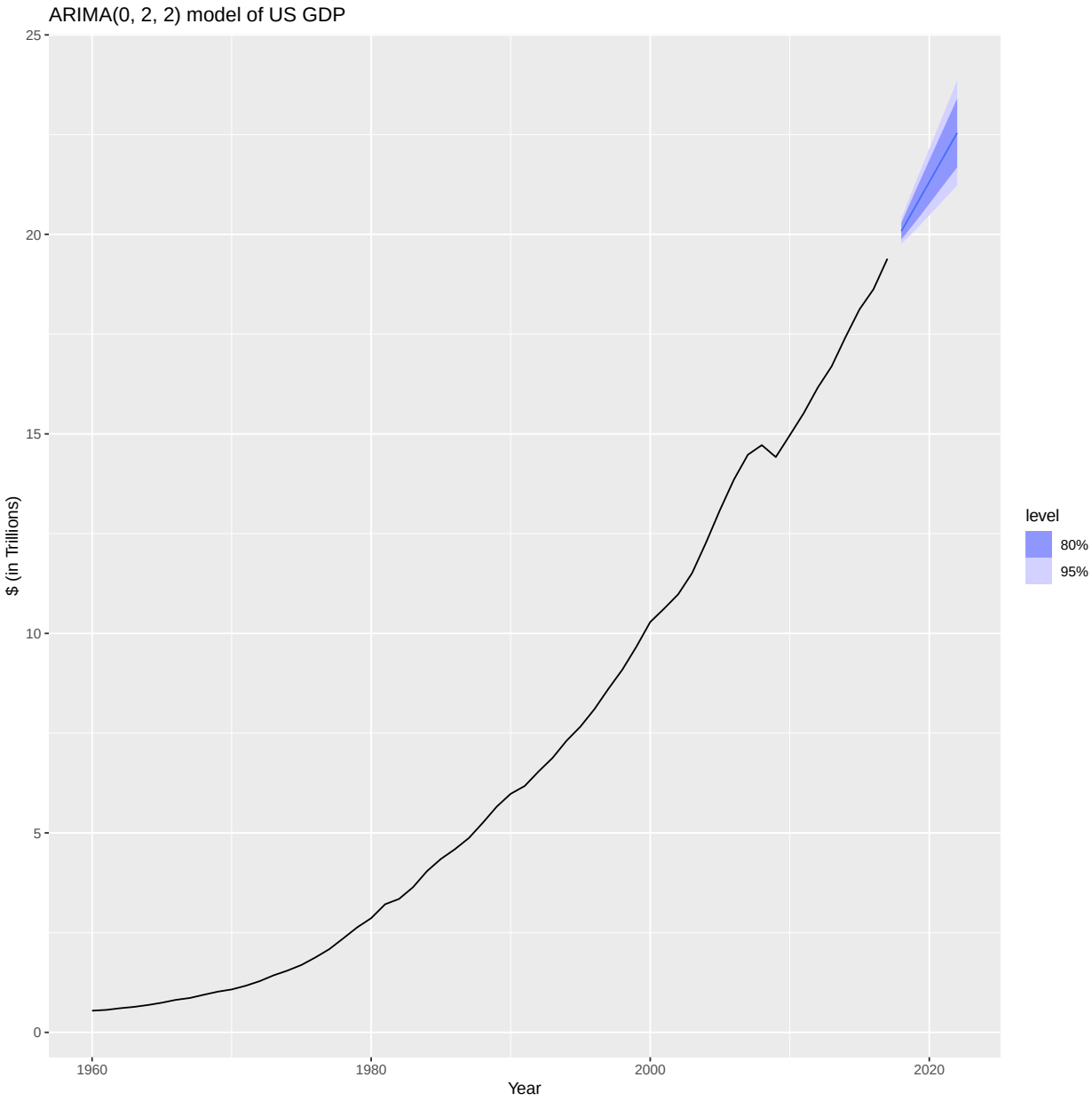
To test different models, I allowed `p`, `d`, and `q` have values up to 3 but excluding the values that the auto generated `ARIMA()` model gave. When doing so, the function returned an ARIMA(1, 2, 1). Comparing the AICc, the auto selected model has a value of `-40.06` while the second model has one of `-39.74` which suggests that these two models perform very similarly to each other although the auto selected one does appear to be better.

4. choose what you think is the best model and check the residual diagnostics;

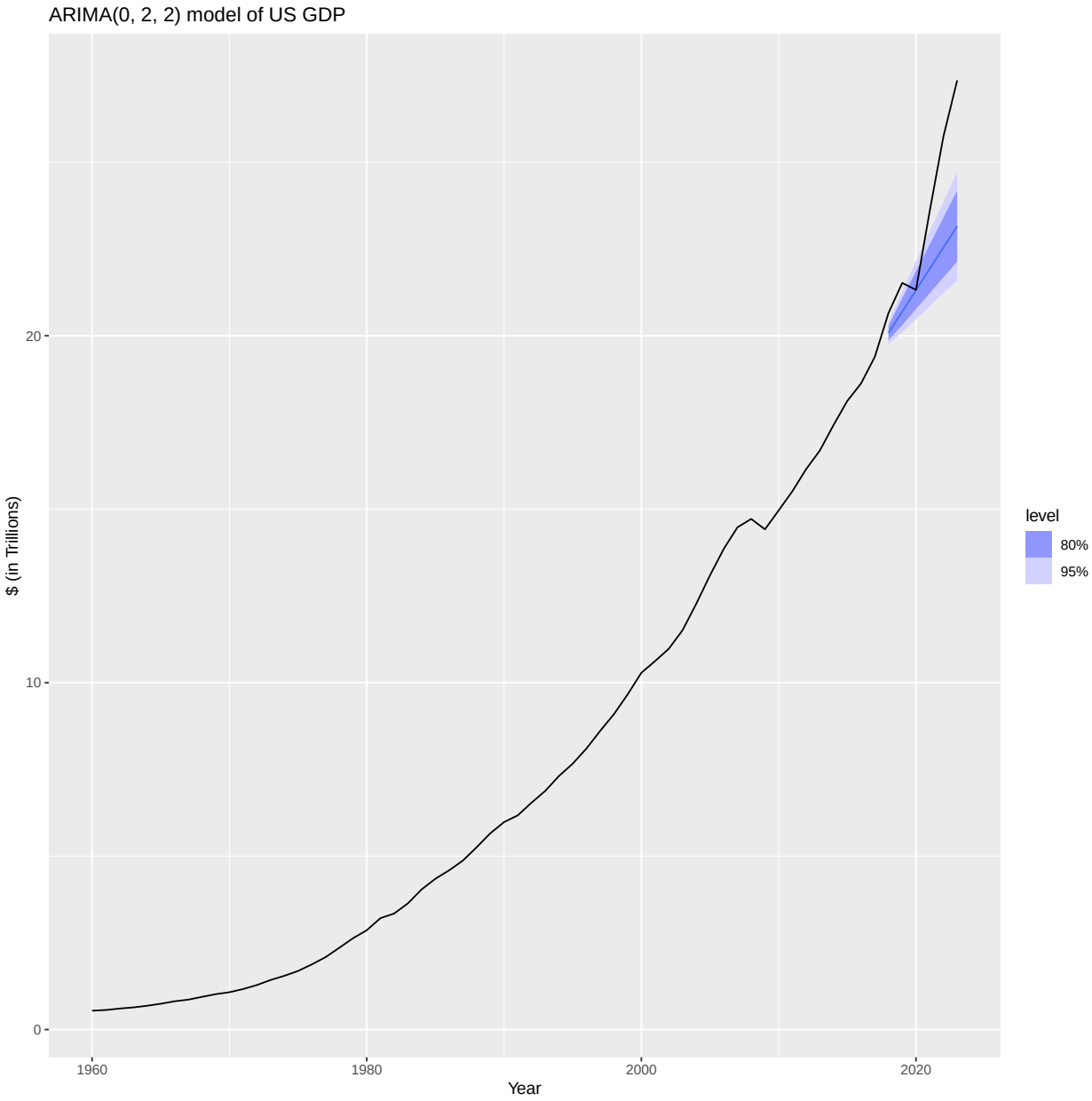


Using the ARIMA(0, 2, 2) model, the plot above shows that all of the autocorrelations are within the threshold limits and that the distribu8tion of residuals are relatively normal. There is one notable outlier which is 2008.

5. produce forecasts of your fitted model. Do the forecasts look reasonable?

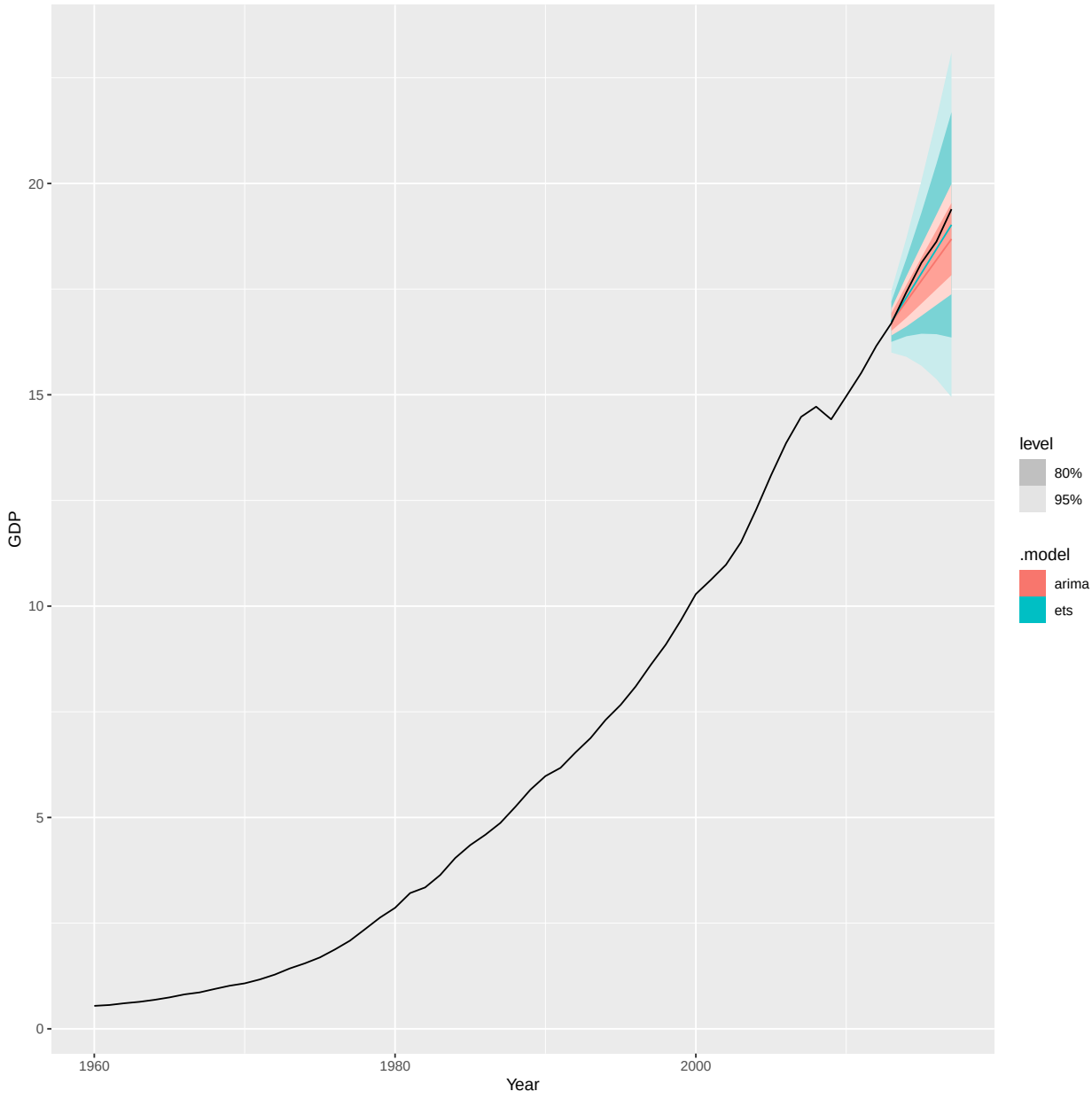


This forecast seems reasonable as it continues the typical trend of growth that the US GDP has had since 1960. Testing it with the latest data from worldbank.org:



Looking at the data updated, it seems that the US GDP's growth has outpaced the forecasted growth generated using the ARIMA(0, 2, 2) model.

6. compare the results with what you would obtain using `ETS()` (with no transformation).




```
## # A tibble: 2 x 10
##   .model .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE   ACF1
##   <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 arima  Test  0.350 0.427 0.360 1.87   1.94 1.16  1.14  0.253
## 2 ets    Test  0.177 0.222 0.191 0.944  1.03 0.613 0.591 0.0812
```

```
## Warning in report.mdl_df(gdp_ets_fit): Model reporting is only supported for
## individual models, so a glance will be shown. To see the report for a specific
## model, use `select()` and `filter()` to identify a single model.
```

```
## # A tibble: 2 x 11
##   .model   sigma2 log_lik   AIC  AICc   BIC     MSE   AMSE     MAE ar_roots
##   <chr>     <dbl>   <dbl> <dbl> <dbl> <dbl>   <dbl>  <dbl>   <dbl> <list>
```

```
## 1 ets    0.000495    29.4 -48.7 -47.4 -38.9  0.0286  0.133  0.0178 <NULL>
## 2 arima  0.0260      21.4 -36.8 -36.3 -31.0 NA      NA     NA      <cpl [0]>
## # i 1 more variable: ma_roots <list>
```

In the above accuracy and model reports, we can see that the MAPE and the AICc of the `ETS()` model are lower, providing evidence that the `ETS()` model is better than the `ARIMA(0, 2, 2)` model. Although the graph shows that the `ARIMA()` model has a much narrower prediction confidence interval and the actual values are within that window.