Assignment 3 Due Friday, October 18 at midnight

Your name

1. Math

Recall the so-called "trend-stationary" model

$$x_t = \mu_t + y_t$$

1. [3 points] There are three terms (symbols) in that equation. Which symbol corresponds to the trend, and which corresponds to the stationary part? What would you call the remaining symbol?

μ_t is the trend, and y_t is the stationary error term as defined in the book. However, depending on what μ_t is, the entire series (x_t) could be stationary as well (for example if $\mu_t = 5$ or some other constant.)

2. We have considered a few possible models for the trend:

$$\mu_{LM,t} = \beta_0 + \beta_1 t$$
$$\mu_{RW,t} = \delta + \mu_{RW,t-1} + w_t$$

3. [2 points] What do "RW" and "LM" stand for?

RW- random walk LM- linear model

3. [4 points] Between $\mu_{LM,t}$ and $\mu_{RW,t}$, which has "more interesting" temporal structure[2 points]? Why[2 points]?

$\mu_{RW,t}$ is "more interesting" because, while the trend structure is linear, it is stochastic (random) because of the w_t term. Also, it has an autoregressive structure. So the random walk is "more interesting".

4. [4 points] Write down the equations for $x_{LM,t}$ and $x_{RW,t}$

This just amounts to plugging in the equation for the trend. For the linear model:

$x_{LM,t} = \mu_{LM,t} + y_t = \beta_0 + \beta_1 + y_t$

For the random walk:

$x_{RW,t} = \mu_{RW,t} + y_t = \delta + \mu_{RW,t-1} + w_t + y_t$

Note that the random walk has two "noise" terms– one in the trend (w_t) and one for the overall model (y_t).

Consider this (edited) excerpt from the textbook (page 49):

Differencing vs. Detrending

One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation. One disadvantage, however, is that differencing does not yield and estimate of the stationary process y_t.
For example, if we difference a random walk x_t with drift δ,

$$x_t - x_{t-1} = \delta + w_t + y_t - y_{t-1}.$$

If an estimate of y_t is essential, then detrending may be more appropriate. This would be the case, for example, if we were interested in the business cycle of commodities.
If the goal is to coerce the data to stationarity, then differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed, that is, when using $\mu_{LM,t}$ as the trend model, we have:

$$x_t - x_{t-1} = \beta_1 + y_t - y_{t-1}$$

Because differencing plays a central role in time series analysis, it receives its own notaiton. The first difference is denoted:

$$\nabla x_t = x_t - x_{t-1}$$

As we have seen, the first difference eliminates a linear trend. A second difference can eliminate a quadratic trend, and so on. Differences of order d are denoted:

$$\nabla^d = (1 - B)^d$$

5. [1 point] Take a look in the book. What is B called?

B is the "backwards shift operator". You can think of it like the `lag` functions in R, in that if you apply the backwards shift operator to a time series x_t, it gives you x_{t-1}, the lagged version of the series. This is then used to define the differencing operator, ∇, which is like the `diff` function in R.

6. [6 points] Why is it an "advantage" that no parameters are estimated in the differencing operation?

Estimating parameters means that estimators (formulas) for your parameter estimates must be derived (figured out using math). So, not having to do that work is an advantage. Also, even if that theory has already been derived, you still need to make sure you have met any assumptions that the formulas rely on and also ensure you have enough data.

7. [6 points] Connect part 2 question 9 below to a sentence in the above excerpt.

Part 2 question 9 asks why you can't detrend a differenced series. You cannot immediately detrend a differenced series because you have not estimated a trend (at least, not without taking additional steps). In order to estimate a trend, you have to estimate parameters, which is mentioned in the first sentence of the excerpt.

7. [10 points] What kind of plot might you make to check if the data has been "coerced to stationarity"?

I would make a time series plot, as I find it easier to see trends that way. However, as we saw when we made acf's of the linear model data from Lecture 7 (the "on your own" activity), you can also sometimes pick up on trend nonstationarity in the acf as well.

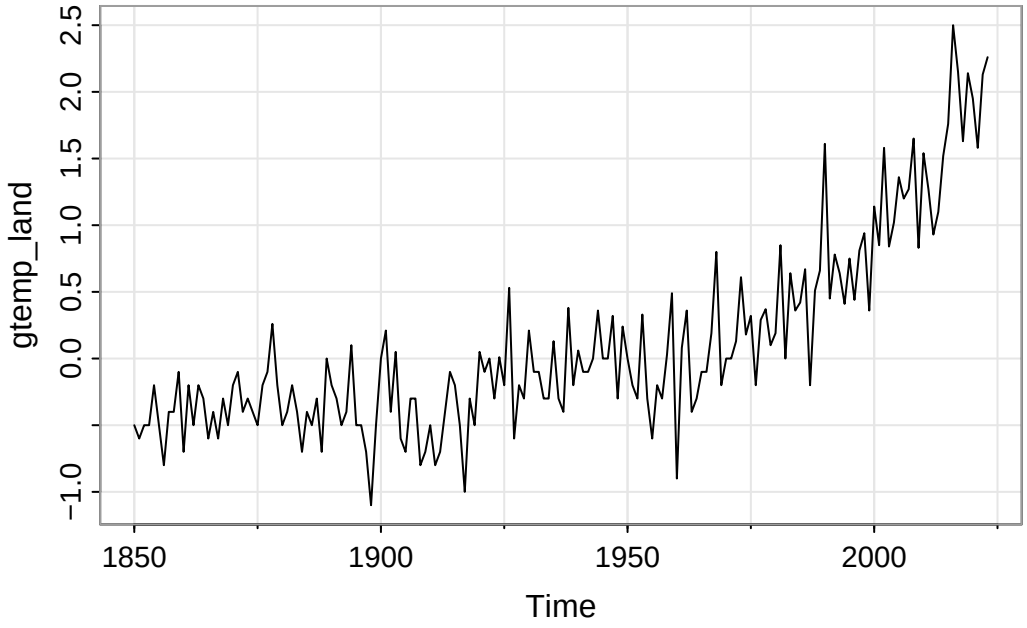
8. [3 points] Rate your math anxiety (1 = effortless, 100 = nightmare) while working on this problem.

I hope this wasn't too bad! It was mostly just figuring out notation, not any computations.

2. Data Analysis (code)

1. [10 points] Adapt the code from Example 1.2 in the book to plot just the global land temperature series in a time series plot. [2 points]. Describe the structure of the trend and/or seasonal components, if present[8 points].

```
library(astsa)
tsplot(gtemp_land)
```



2. [4 points] How frequently were the observations collected?

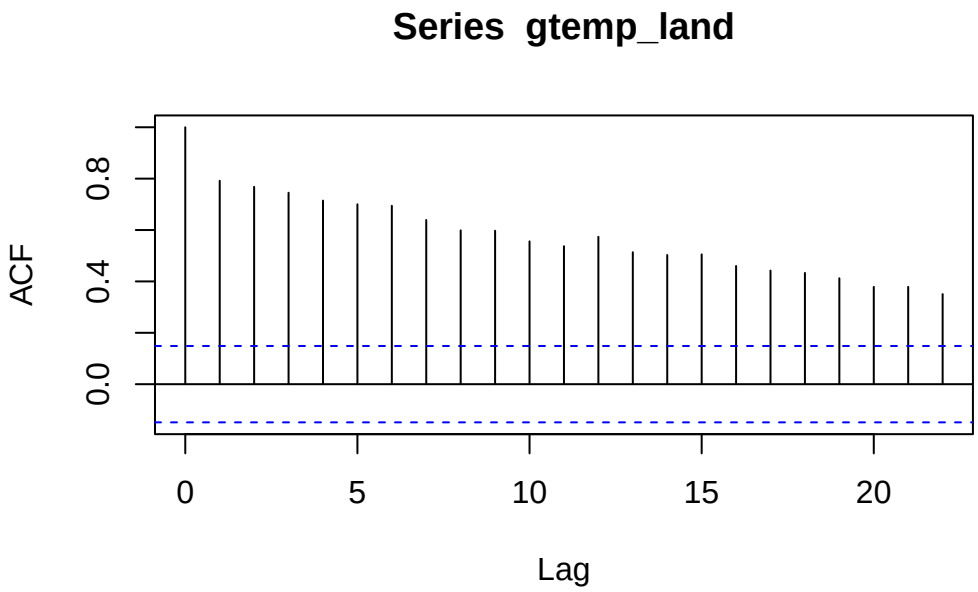
```
diff(time(gtemp_land))
```

```
Time Series:
Start = 1851
End = 2023
Frequency = 1
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Yearly.

3. [5 points] Plot the autocorrelation function of the global land temperature series. Comment on the temporal structure.

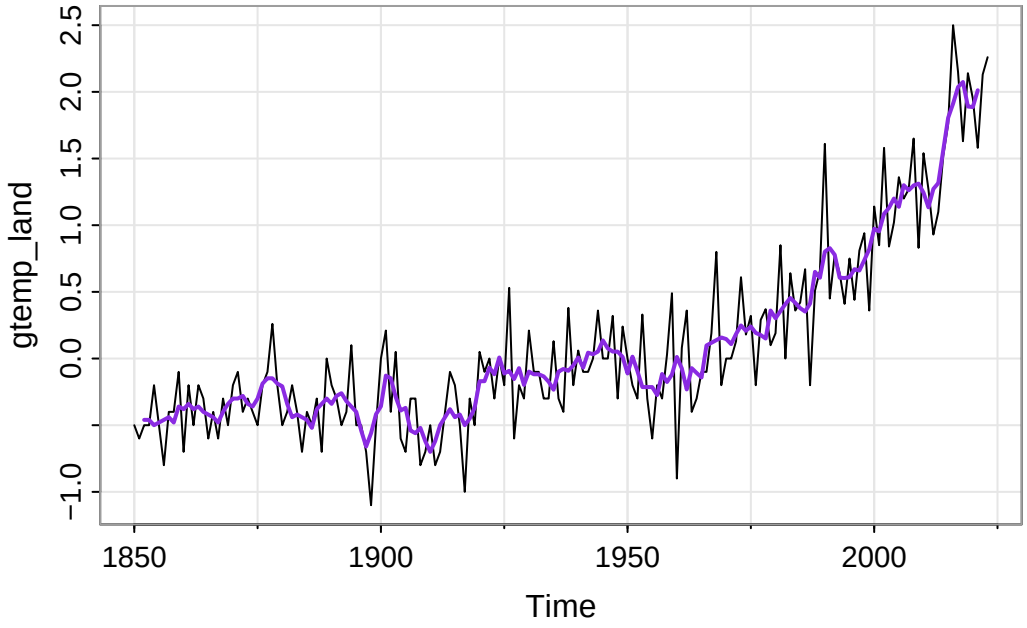
```
acf(gtemp_land)
```



There appears to be an increasing trend, and this appears to be increasing faster than linear (slope is steeper the further in the series you go).

4. [15 points] Estimate the trend of the series using a symmetric, equally weighted 5-point moving average[4 points]. Plot the trend estimate on top of the data [4 points]. Comment on the trend– does it reveal any patterns difficult to see in the data?[7 points]

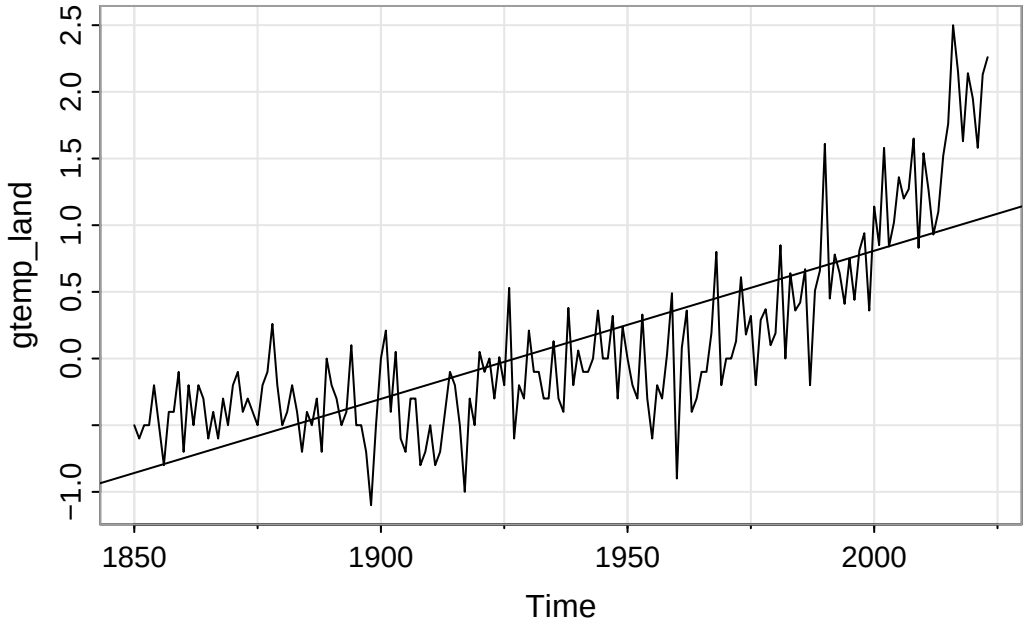
```
ma_trend_estimate <- stats::filter(gtemp_land, filter = rep(1/5,5), method = "convolution")
tsplot(gtemp_land)
lines(ma_trend_estimate, col = "blueviolet", lwd = 2)
```



No, the trend does not appear to pick up any additional structure not in the data. The increasing trend is still present, but there do not appear to be any cycles/pseudo-cycles (compare this to the MA estimate for the SOI data, where the moving average revealed a 4-5 year pattern).

5. [10 points] Estimate the trend of the series using a linear regression on time[3 points]. Plot the trend estimate on top of the data[3 points]. Is this trend estimate comparable to the moving average?[4 points]

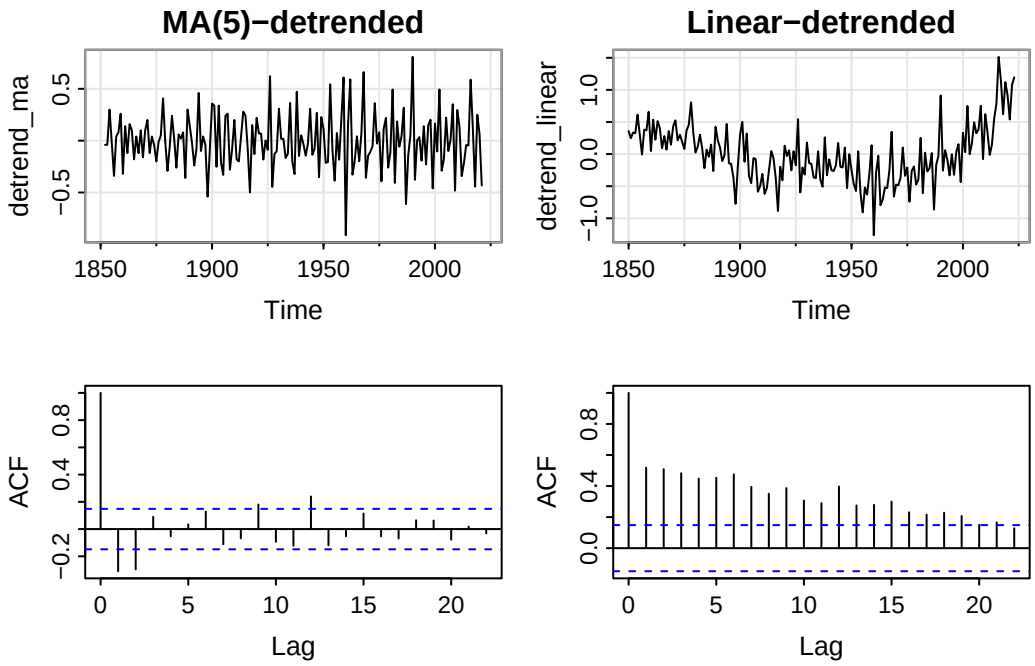
```
linear_trend_estimate <- lm(gtemp_land ~time(gtemp_land))
tsplot(gtemp_land)
abline(linear_trend_estimate)
```



No. This trend does not track the pattern of the global temperature anomaly series well at all, compared to the moving average.

6. [15 points] De-trend the data with respect to each of the trends you estimated [4 points each]. Make a time series plot of each result [1 point each] and an acf of each result [1 point each]. Comment on the temporal structure in each of the plots.

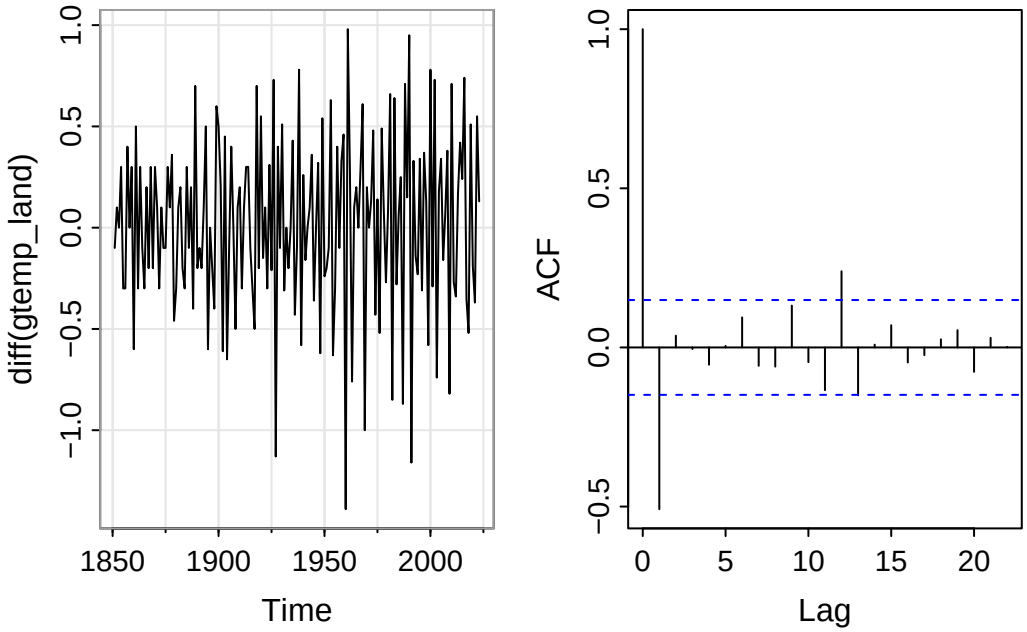
```
detrend_ma <- gtemp_land - ma_trend_estimate
detrend_linear <- gtemp_land - fitted.values(linear_trend_estimate)

par(mfrow = c(2,2))
tsplot(detrend_ma, main = "MA(5)-detrended")
tsplot(detrend_linear, main = "Linear-detrended")
acf(detrend_ma, na.action = na.pass)
acf(detrend_linear)
```



The MA(5)-detrended series looks very similar to white noise, though there may be a little bit of temporal structure left for small lags, though this could just be random fluctuations. The linearly detrended data has a curved pattern, and the acf reveals quite a lot of residual temporal structure (slowly decreasing pattern in the acf).

7. [20 points] Difference the time series and plot the result[5 points]. Also compute the acf and plot it[5 points]. Comment on whether the differencing has "coerced the data to stationarity"[10 points].

```
par(mfrow = c(1,2))
tsplot(diff(gtemp_land))
acf(diff(gtemp_land))
```



The data does look stationary (looking at the time series plot). We do not see any trend non-stationarity suggested by the acf plot either, though there appears to be some lag-1 temporal structure.

8. [11 points] Estimate the sum of squared error for both the linear regression and the moving average [3 points each]. What do they suggest about which model is "better"?[3 points] Does this agree with your visual assessments in 3 and 4?[2 points]

```
sum(detrend_ma^2, na.rm = T)
```

```
[1] 12.42912
```

```
sum(detrend_linear^2)
```

```
[1] 34.93531
```

The sum of squared errors for a given trend estimate is the sum of the squared residuals. The detrended series are in fact the residuals, so we can just square and add them up (ignoring the missing values in the MA-detrended series since they are just a result of estimating the moving average, rather than true missing data).

The SSE is smaller for the MA, which means that the errors for the moving average trend estimate are smaller than for the linear trend. This agrees with our visual assessment that the moving average tracks the data better.

9. [10 points] Why can't you do part 6 for the differenced series?

We cannot de-trend after differencing because differencing does not estimate a trend.

10. [3 points] Ensure that when you render the document to .html to turn in that you have `embed-resources: true` in the options at the top of the document (you may also turn in a pdf). Also, set message to false for all the code chunks.

:)

3. The literature

The original paper for the data set you just analyzed can be found here.

1. [2 points] Set a timer for 5 minutes. Find one sentence you feel you understand, and one you do not understand but would like to understand better.

:)

2. [4 points] Take a look at Figure 6. There are six time series plotted. One is the data plotted in number 1 of the coding portion, the other is the moving average estimate estimated in number 3. Which of the six series are they?

We are plotting the global temperature series, so it would be the top plot that is labeled "Global". I'm not actually sure if this directly corresponds to only land or only sea, or some combined amount, but we are not splitting based on hemispheres as they are in the paper.

3. [4 points] The textbook for our class states on page 3 that "the data are annual temperature anomalies averaged over the Earth's land area.". Does the book state what is specifically meant by "anomaly" here?

Yes, it is the difference from the 1991-2020 average.

4. [2 points] Consider the concept of averaging over the Earth's land area, then take a look at Figure 1. Describe in terms of the "circles" how you might calculate that global average.

Each circle corresponds to a station, so we would just average the values collected at each station. We could also do something more complicated and use the circles to define a spatial surface, then integrate over that spatial surface to find the average (spatial statistics!!).

5. [15 points] Figure 3 plots correlation coefficients between annual mean temperature changes for pairs of randomly selected stations having at least 50 common years in their records. How can you rephrase "annual mean tempearture changes" in terms of detrending or differencing?

Each dot on the plot corresponds the correlation between two time series–a pair of stations. They say that these series are "annual mean temperature changes", which would imply differencing (subtract this year's value from last year's value to get the change).

4. Weights

1. [6 points] One might consider the points I have given to each numbered problem as "weights". Based on the weights, which content is the most important? Answer on both the individual problem level and the section level (i.e. math, code, etc).

Math-