Statistics 5350/7110
Forecasting

Lecture 8
Detrending

a.k.a., Manufacturing Stationarity

Professor Robert Stine

Admin Issues

- Questions
  - No office hours today
- Assignments
  - Assignment 2
  - Downloading data files from Canvas (in Assignments folder)
- Quick review
  - Multiple regression models for time series
  - Diagnostic plots
  - Seasonal patterns and spurious correlation
  - Software: dynlm for building regression models for time series in R

Today's Topics

Text, §3.2

- Time series regression
  - Finish regression modeling example from Lecture_7.Rmd
  - Comparing lm to dynlm
- Detrending a time series
  - Many procedures in forecasting presume stationarity (e.g. estimating autocovariances)
- Question: How to detrend?
  - Is there a deterministic trend (use regression) or is it a random walk (difference the data)
  - Does the choice matter? Yes!
  - Tradeoffs if we make the wrong choice
- Examples feature climate data
  - Global temperature

Finding the Stationary Process

- Motivating model
  - Observed time series is the sum of a mean plus a zero-mean stationary process

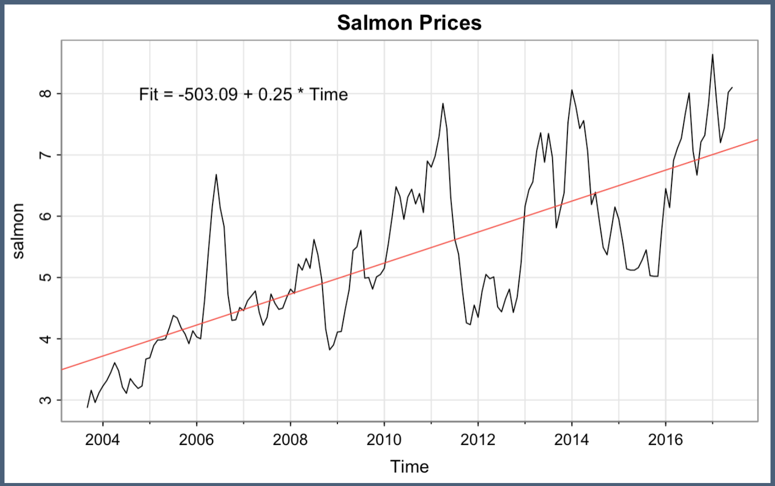
    $$X_t = \mu_t + Y_t \qquad \text{or simpler} \qquad X_t = \mu_t + w_t$$

    where $E(X_t) = \mu_t$, $\{Y_t\}$ is a stationary process, and $\{w_t\}$ is white noise.
  - Analyze the otherwise hidden stationary process.
  - Forecast: Extrapolate mean as "trend" and predict stationary process
- Obvious choices to obtain stationarity
  - Estimate a deterministic trend and subtract it from the data; analyze the residuals.
  - Difference the data to remove a random walk.
- What could go wrong?
  - Consider taking the wrong action
  - Difference the data when the trend is non-stochastic
  - Fit a deterministic trend when μ_t is a random walk

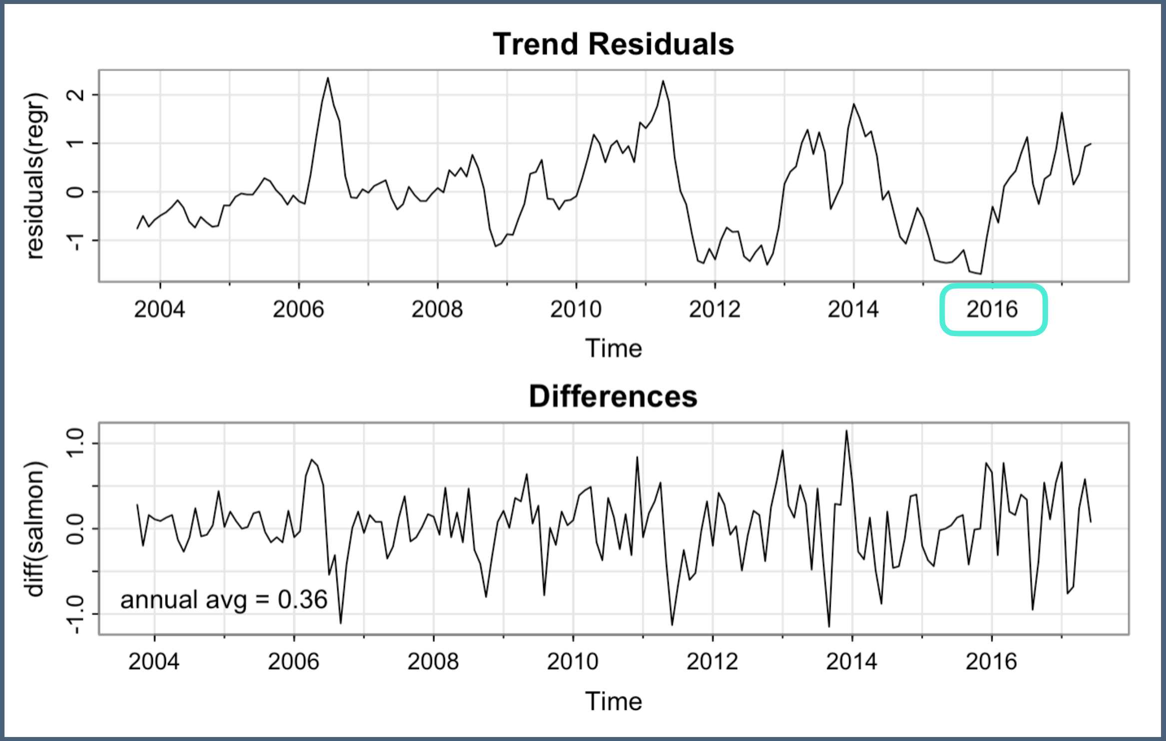
We'll do this modeling simultaneously later

Does the choice matter?

- Yes!

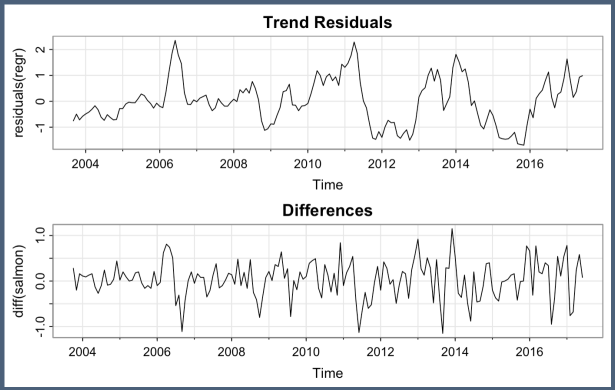


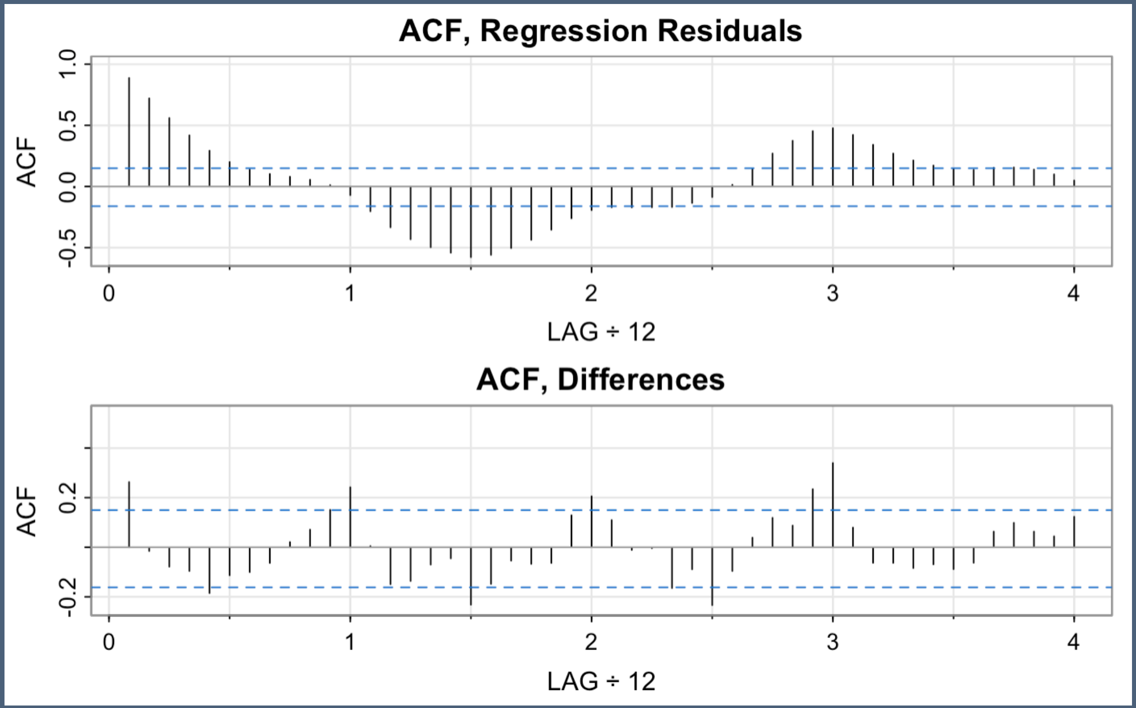
dynlm



Does the choice matter?

- Yes!
  - Long term dependence in the residuals
  - Annual pattern in differences





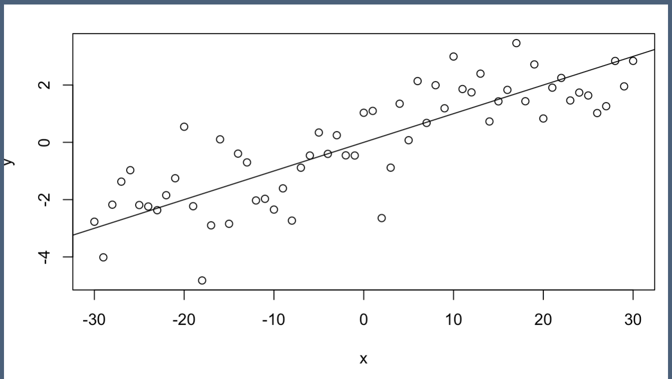
Fitting a Deterministic Trend

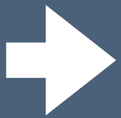
$X_t = \mu_t + Y_t$

- Suppose the mean is deterministic
  - Example: a linear trend is correct model $\mu_t = \alpha + \beta\, t$
  - Estimate the model and subsequently work with residuals $\widehat{Y}_t = X_t - \hat{\mu}_t = Y_t + (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})\, t$
  - Some of the trend "leaks" into the residuals (assuming this model is correct)
  - If we care about β, then regression gives a very precise estimate (presuming assumptions)
- Simulation comparison
  - Simulate data with a trend
  - Estimates have same mean, but slope is less variable

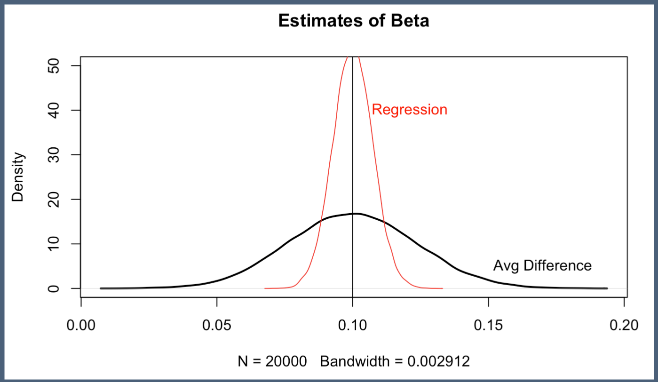
Details in Rmd file

Var(avg diff) ≈ (n/6) Var(b)





simulate
20000
samples



Fitting a Deterministic Trend

$X_t = \mu_t + Y_t$

- Suppose the mean is deterministic
  - Example: a linear trend is correct model $\mu_t = \alpha + \beta\, t$
  - Estimate the model and subsequently work with residuals $\widehat{Y}_t = X_t - \hat{\mu}_t = Y_t + (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})\, t$
  - Regression estimate of β is much more efficient than differencing (smaller SE by factor 1/√n)

- Suppose we got it wrong
  - We model as a trend but the mean is a random walk… $\mu_t = \delta + \mu_{t-1} + w_t$
  - What happens if we fit a trend when the data is a random walk?

$$\widehat{Y}_t = X_t - \hat{\mu}_t = Y_t + \left(\mu_t - \hat{\alpha} - \hat{\beta}\, t\right)$$

  Residuals mix the stationary process Y_t with deviations of fitted trend from a random walk…
  - Not a stationary process!
  - Plus, our precise claims about the slope are wrong (LS regression inflates the precision)

Differencing

- Special notation for the backshift operator B
  - Define operator B as a time shift $B\,X_t = X_{t-1}$
  - Differencing in terms of B $\nabla\,X_t = X_t - X_{t-1} = (1 - B)\,X_t$
- Powerful notation
  - Treat the operator B as an algebraic symbol, as it if represents a number
  - Second differences

$$\nabla^2\,X_t = (1 - B)^2\,X_t = (1 - 2B + B^2)X_t = X_t - 2X_{t-1} + X_{t-2}$$

  - Differences of AR(1)

$$X_t = \phi X_{t-1} + w_t \;\Rightarrow\; (1 - \phi B)X_t = w_t \;\Rightarrow\; X_t = \frac{1}{1 - \phi B} w_t$$

  If we assume that |ϕ| < 1 and treat B as if |B| = 1 then

$$\frac{1}{1 - \phi B} w_t = \left(1 + \phi B + (\phi B)^2 + \cdots\right) w_t = w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \cdots$$

  - Much more of this to come in our analysis of ARIMA models

This notation was popularized by Box and Jenkins in an influential book on time series analysis.

Differencing

$X_t = \mu_t + Y_t$

- Suppose the mean function is a random walk
  - Mean function has possible drift $\mu_t = \delta + \mu_{t-1} + w_t$
  - Differencing leaves a stationary process with "no estimation" needed

$$\nabla X_t = X_t - X_{t-1} = \mu_t + Y_t - (\mu_{t-1} + Y_{t-1}) = \delta + w_t + (Y_t - Y_{t-1}) = \delta + w_t + \nabla Y_t$$

  - Differences ∇X_t form a stationary process (eqn 3.24)

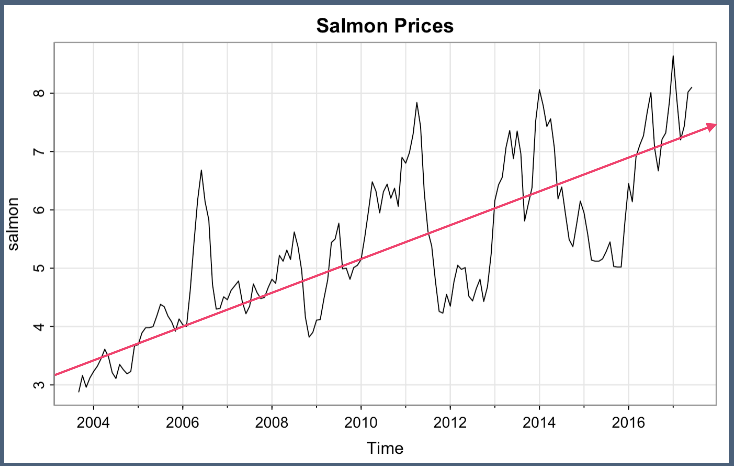
  If $\{Y_t\}$ is a stationary process, then $\{\nabla Y_t\}$ is a stationary process. If $u_t = y_t - y_{t-1}$, then

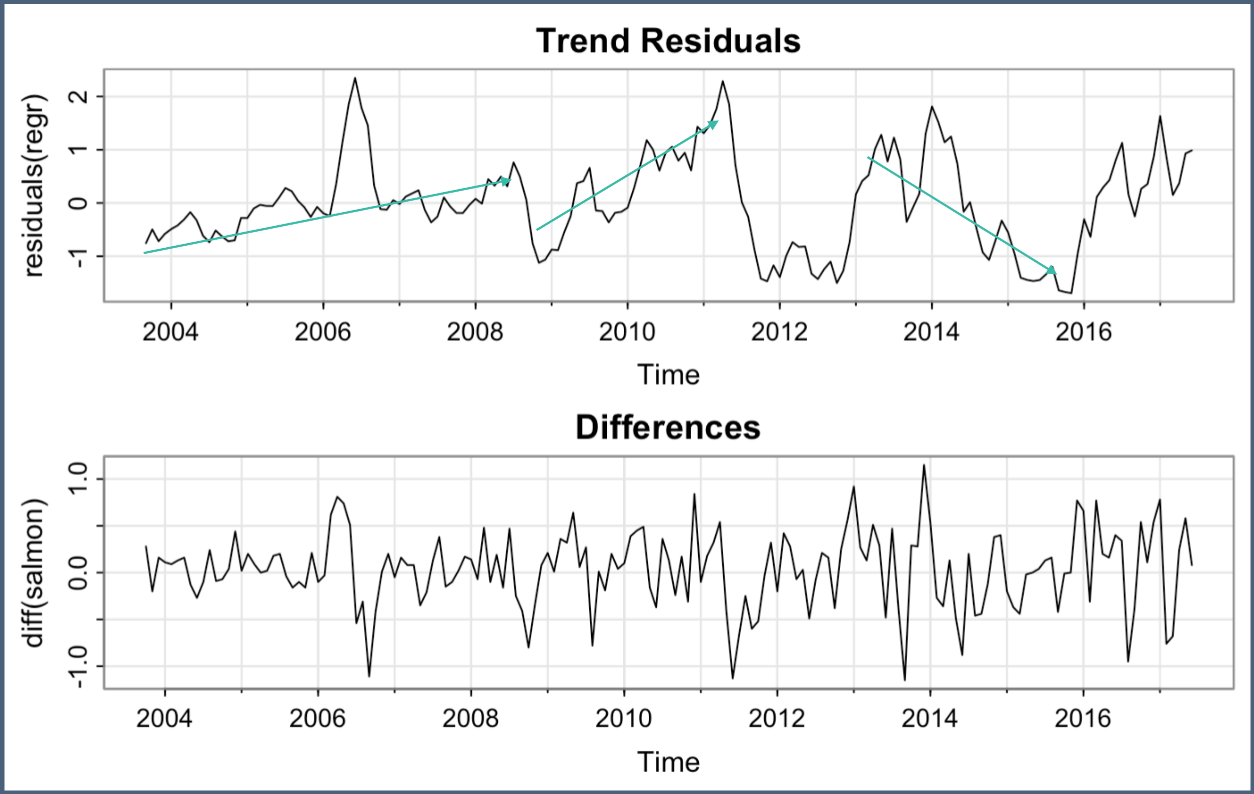
$$\text{Cov}(U_{t+h}, U_t) = \text{Cov}(Y_{t+h} - Y_{t+h-1}, Y_t - Y_{t-1}) = 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1)$$

  - Hence, we obtain a stationary process, but it's not Y_t.
- Suppose we difference when the mean is a deterministic linear trend
  - The differences are then $\nabla X_t = \nabla(\alpha + \beta t + Y_t) = \beta + \nabla Y_t$
  - We again don't directly observe Y_t, but we again get a stationary process.
- Differencing is a more reliable means to obtaining a stationary process
  - At the cost of a less precise estimate of β than regression when μ_t is deterministic

Back to the example…

- Does this time series appear to have a linear trend?

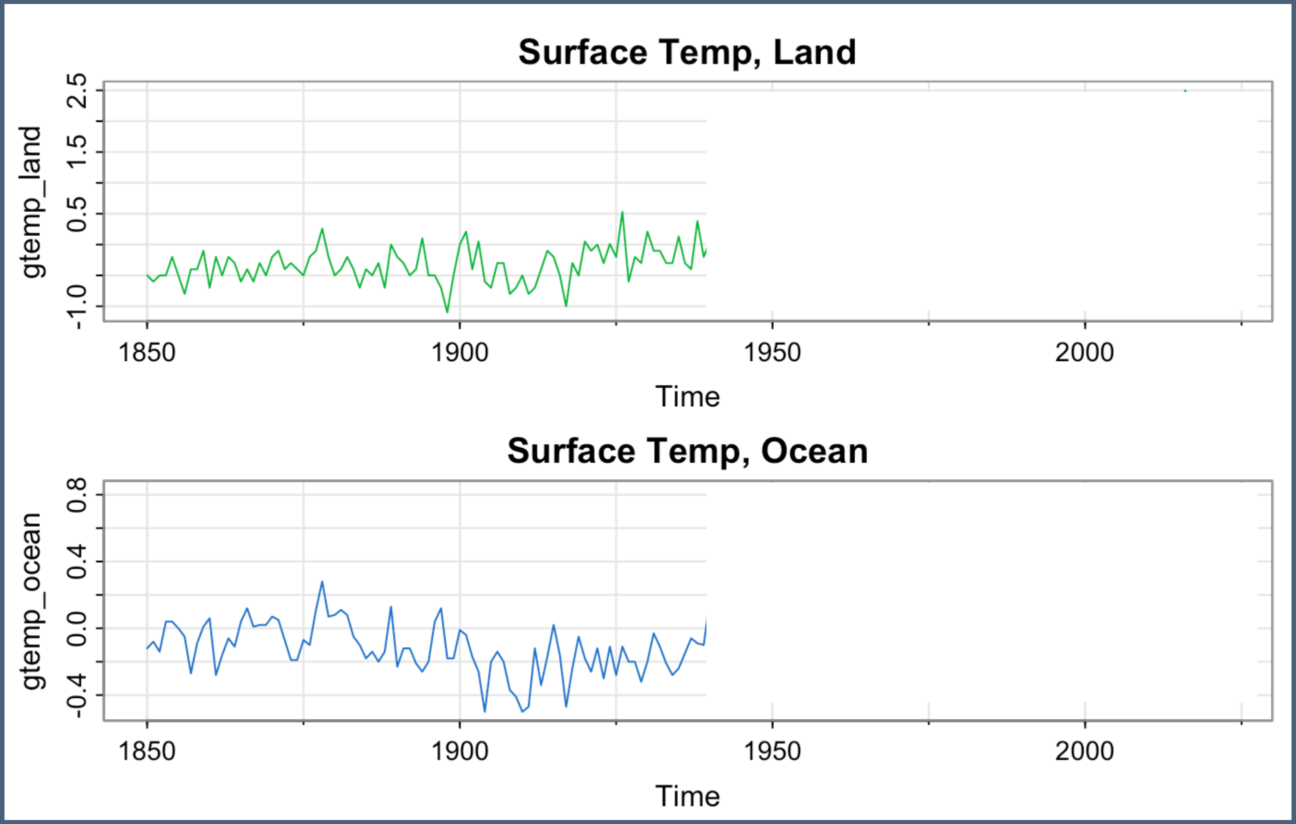




Return to this time series later in course, but differencing looks like the more reasonable approach

Second Example: Global Temperature

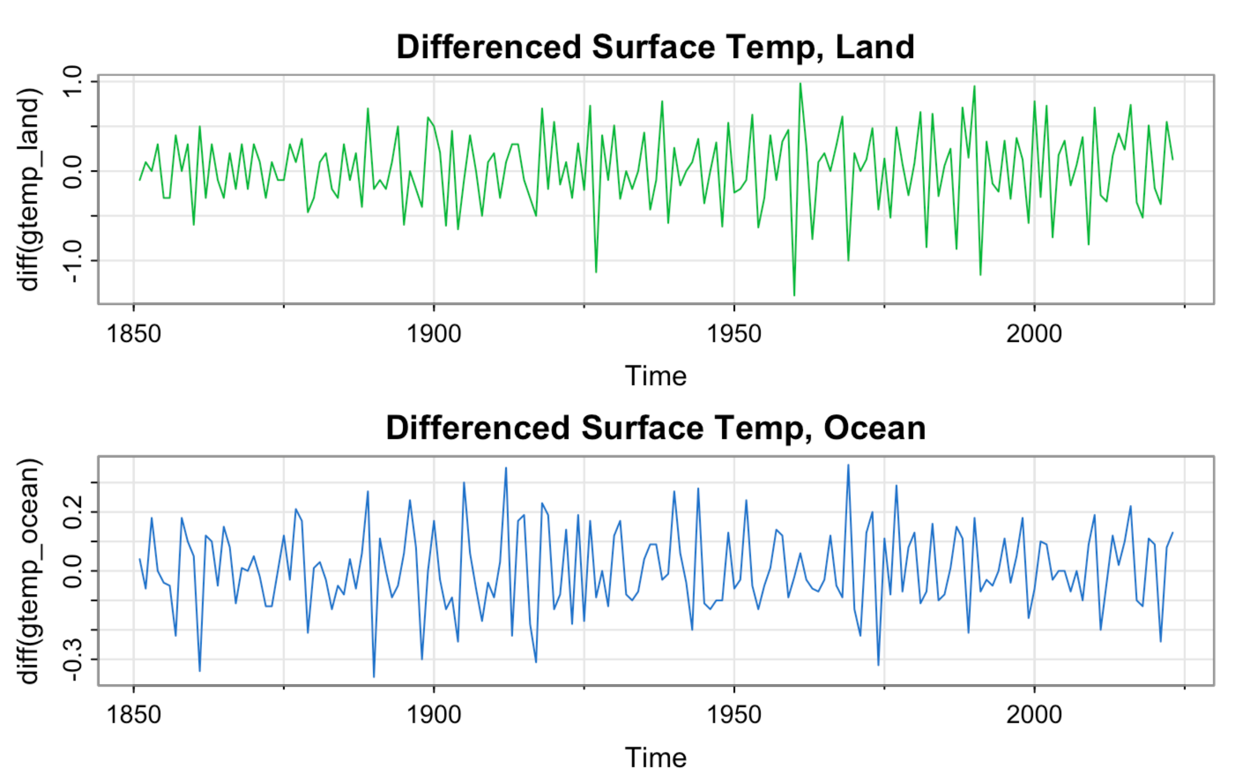
- Global surface temperature deviations
  - Both appear stationary until post WWII economic expansion around 1945-1950.



Variability changes with changes in how these data are obtained.

Differenced Temperature

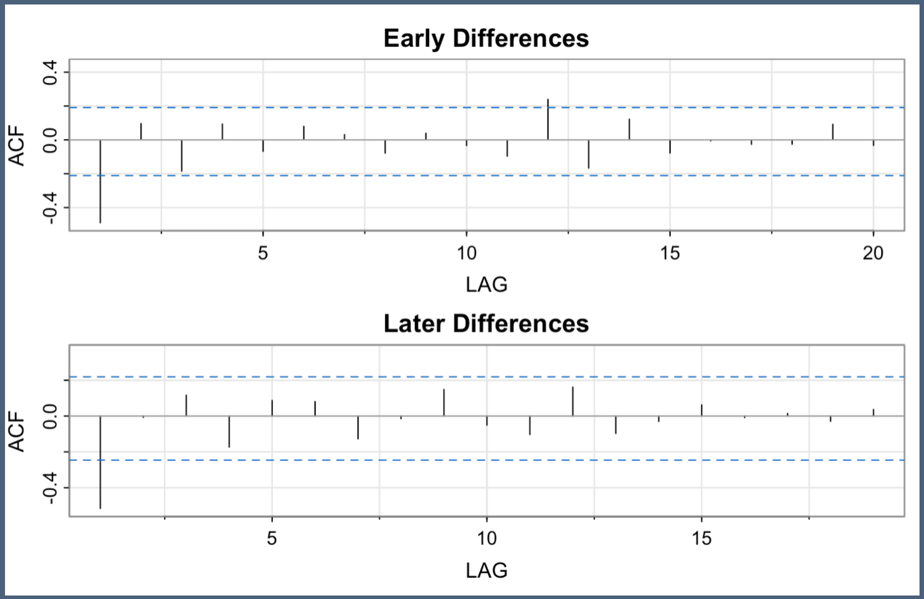
- Differences appear stationary
  - Changes in the drift are not very apparent
  - Changes in variation noticeable in the land temperatures



Huge literature on detecting a "change point" in a time series.

Autocorrelation in Differences

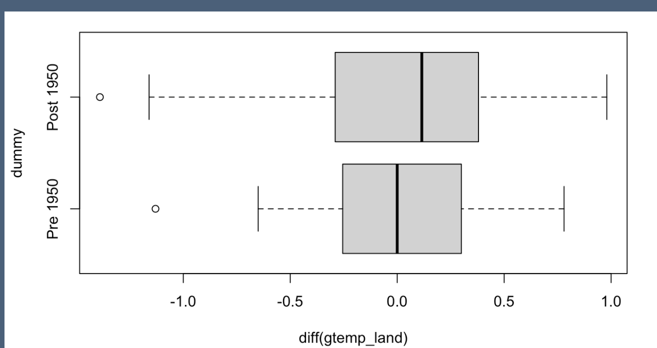
- Differences have autocorrelation
  - Structure of these autocorrelations is consistent with a simple model
  - Suppose $X_t = \mu_t + u_t$ where $\mu_t = \mu_{t-1} + w_t$ is a random walk and u_t is independent white noise.
  - Then the differences are $\nabla X_t = w_t + (u_t - u_{t-1})$
  - Autocovariances of the differences are then $\gamma(1) = -\sigma_u^2$ and $\gamma(h) = 0$ for 1 < |h|.



Similar ACF pre/post 1950.

Comparison of Differences

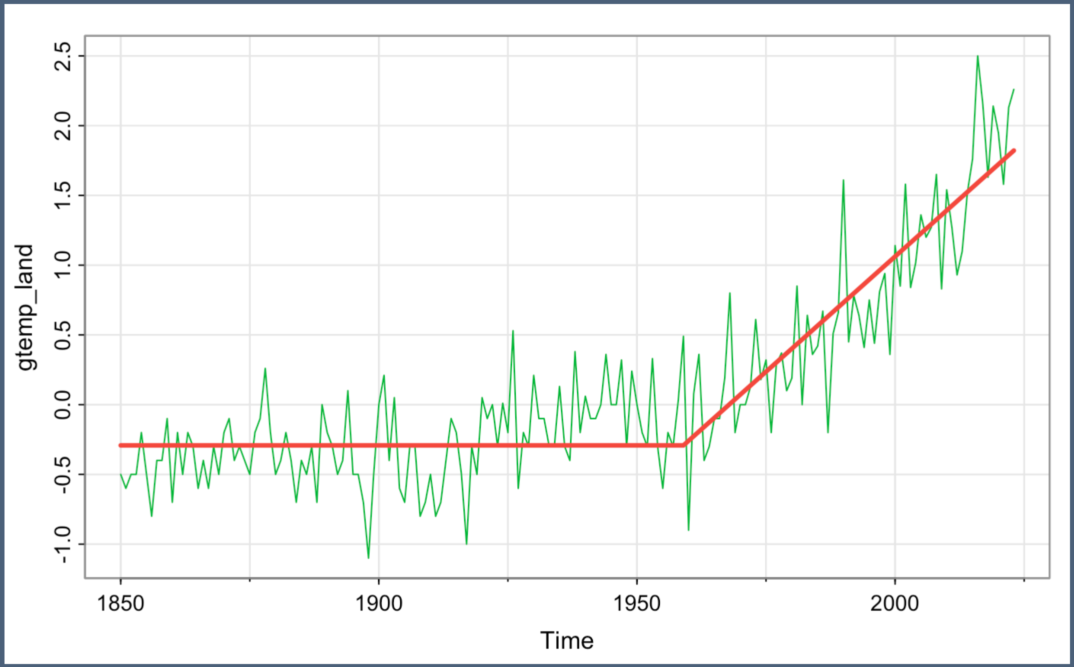
- Two-sample t-test
  - Test null hypothesis that means are same in two time periods
  - As if the observations of the differences were independent (we know they aren't)
- Results
  - Difference of average differences is about 0.02 degrees
  - Nowhere close to significance




```
t = 0.28394, df = 125.2, p-value = 0.7769
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1183409  0.1579860
sample estimates:
  mean of x   mean of y
0.027297297 0.007474747
```

Different Analysis

- Fit a regression to the temperatures
  - Measure the slope since visually chosen change point
  - Use the usual regression test statistics: find larger 0.03 for growth post 1959
- Statistically significant?



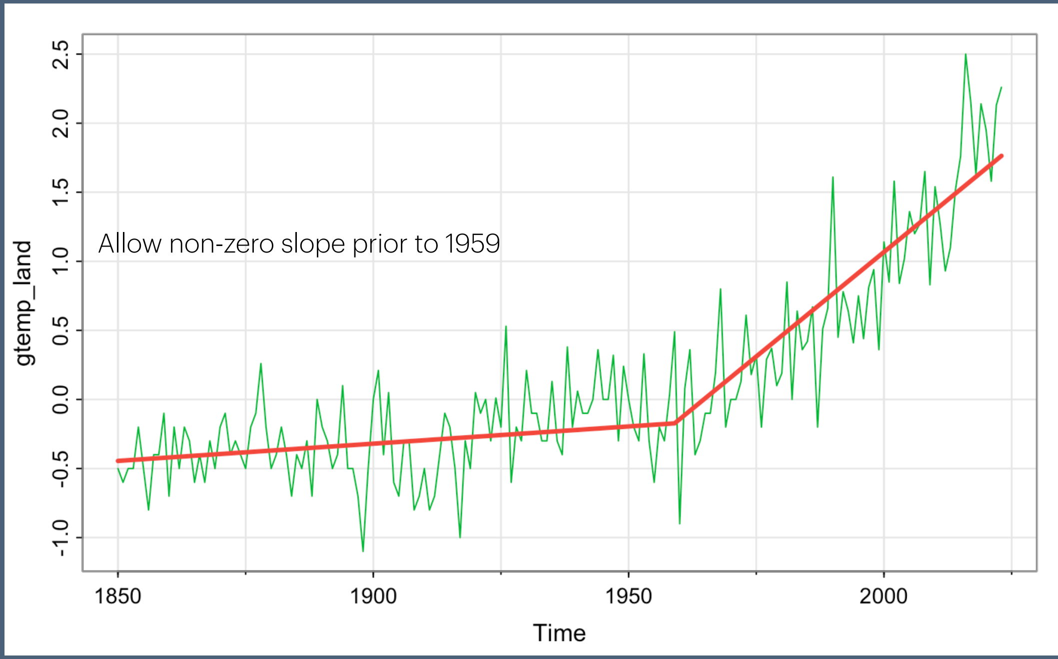

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.292021   0.029418  -9.927   <2e-16 ***
recent_time  0.033025   0.001298  25.452   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3297 on 172 degrees of freedom
Multiple R-squared:  0.7902,    Adjusted R-squared:  0.789
F-statistic: 647.8 on 1 and 172 DF,  p-value: < 2.2e-16
```

This model forces continuous fit

Different Analysis, Enhanced

- Fit a regression to the temperatures
  - Measure the change in the slope at the visually chosen change point
  - Use the usual regression test statistics: again find about 0.03 for growth post 1959
- Statistically significant?



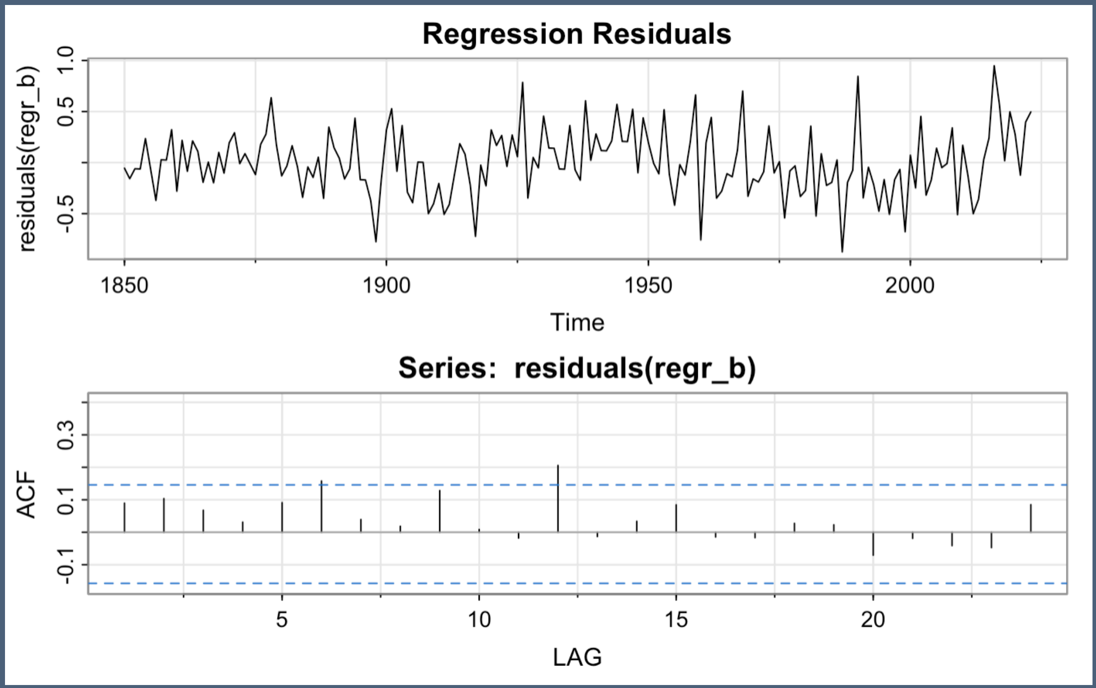

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -5.0619173  1.5823059  -3.199  0.00164 **
time(gtemp_land)  0.0024957  0.0008277   3.015  0.00296 **
recent_time       0.0277596  0.0021582  12.863  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3222 on 171 degrees of freedom
Multiple R-squared:  0.8008,    Adjusted R-squared:  0.7985
F-statistic: 343.7 on 2 and 171 DF,  p-value: < 2.2e-16
```

This model forces continuous fit and allows nonzero slope in the initial data

Residual Analysis

- Not much residual autocorrelation
  - Certainly not like residuals from a random walk.
- Implications?

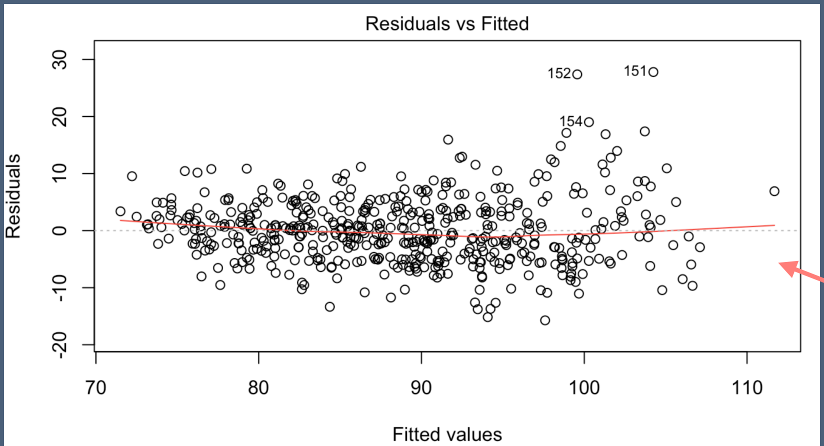


What's next?

Textbook §3.3

- Smoothing data
  - Estimating a mean function non-parametrically
  - Shown in many diagnostic plots: regression diagnostics, scatterplot matrix
  - Applications in cross-sectional vs. time series

Example of a calibration plot
Residuals on Fitted values



What's that curve?