Statistics 5350/7110
Forecasting

Lecture 7
Regression with Time Series Data

Professor Stine

Admin Issues

- Questions
  - No office hours for me on Thursday

- Assignments
  - A1 being graded.
  - Next assignment given out on Thursday

- Quick review
  - Relevance of normality of model errors — why checking QQ plot is important in forecasting
  - Building a multiple regression in R
  - Leverage diagnostics
  - Added variable/partial regression plot

Today's Topics

- More cross-sectional regression
  - Categorical variables
- Regression in the context of time series
  - Response and predictors are time series
- Autocorrelation
  - Residuals imply model errors are not independent
  - Consequences for inference
  - Durbin-Watson statistic
- Residual plots for time series
  - Sequence plot, lag plot, ACF

Finish example from Lecture_6.Rmd

Multiple Regression Model

- Model combines an equation with assumptions for deviations
  - Model for a time series X_t with q predictors $Z_1, Z_2, \dots, Z_q$

$$X_t = \beta_0 + \beta_1 Z_{t,1} + \cdots + \beta_q Z_{t,q} + w_t$$

  - White noise errors: mean zero, common variance σ_w^2
- Is the model well-specified?
  - Does the model make substantive sense? (e.g. Have you mixed stock and flow variables?)
  - Does the model have the relevant explanatory variables?
  - Is the anticipated accuracy good enough?
- Diagnostic issues
  - Do the model errors appear to be independent? (at least uncorrelated!)
  - Do the model effects appear linear? Are fitted values well-calibrated?
  - Are any observations unusually leveraged/influential?
  - If require prediction intervals, are the model errors normally distributed with comparable variances?

Choices of symbols for response (X) and predictors (Z) match the textbook choices.

Residual Autocorrelation

- Problem: Underlying model errors are autocorrelated
  - Common reason: missing features from model, lacking transformation
  - Common reason: wrong lags, missing lags
- Consequences
  - Inference is misleading: Estimates appear significant but in fact are not
  - Prediction isn't as accurate as it could be
- Explanation
  - Easier to quantify consequences of autocorrelation if we consider estimating the sample mean
  - Suppose data are autocorrelated, $X_t = \phi X_{t-1} + w_t$ so that corr$(X_t, X_{t-1}) = \phi$
    But we compute the standard error as if the data are independent.

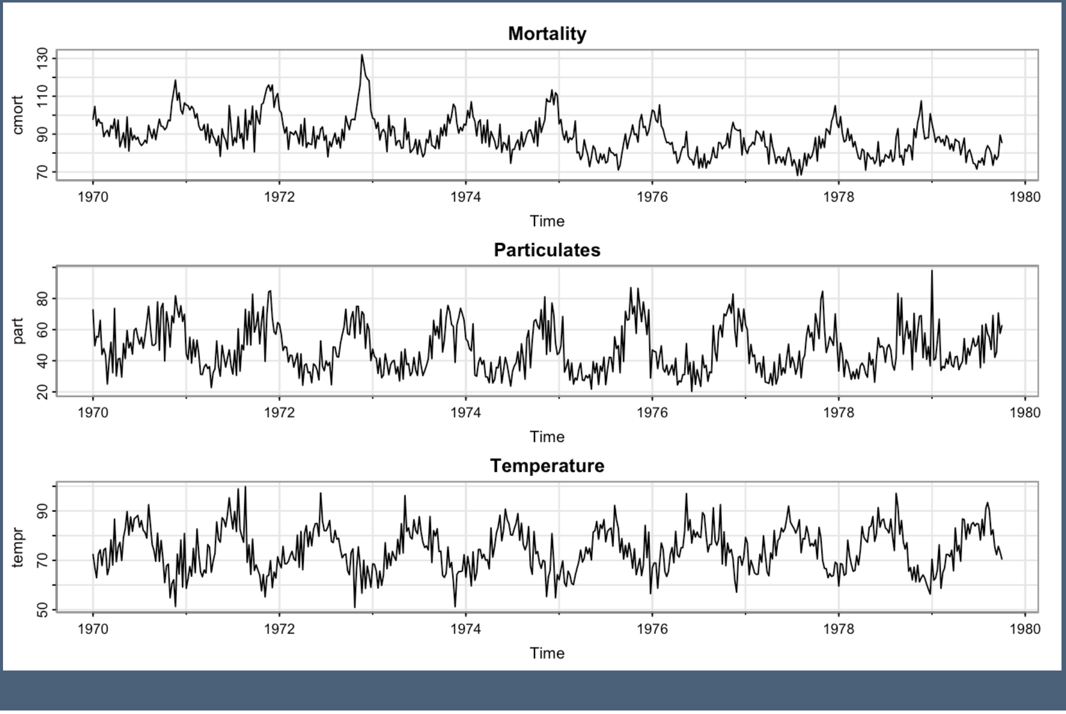
    Independent: Var$(\overline{X}) = \dfrac{\sigma_X^2}{n}$ Correlated: Var$(\overline{X}) = \dfrac{\sigma_X^2}{n} \sum_{h=-n}^{n} \left(1 - \dfrac{|h|}{n}\right) \phi^{|h|} \approx \dfrac{\sigma_X^2}{n} \boxed{\dfrac{1+\phi}{1-\phi}}$
  - Since typically $0 < \phi$, our standard error is too small

Example for Discussion

- Pollution and mortality
  - 10 years of weekly data
  - Los Angeles County
  - CV mortality is the response, with temperature and pollution as explanatory features
- Comments
  - Downward trend in mortality
  - Highly seasonal, with evident annual cycle
  - Series are roughly aligned, with particulates leading mortality
  - Positive correlation with particulates, negative with temperature.

Example 3.5, p 41

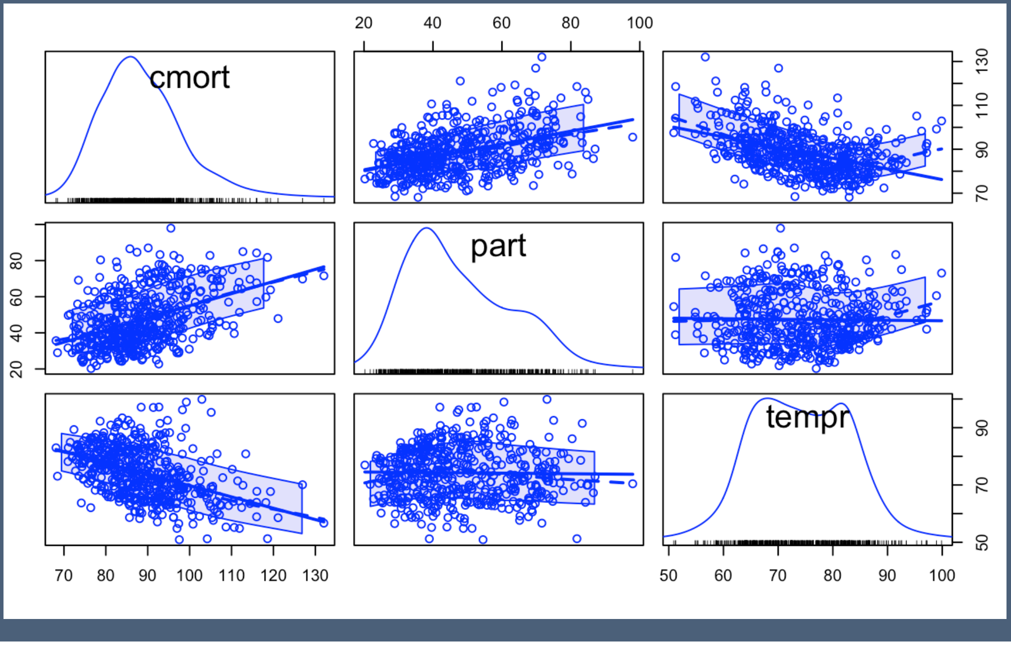
Example 5.16, p 123



Note that mortality rate data is smoothed daily rates... Read the help information for this variable.

Associations

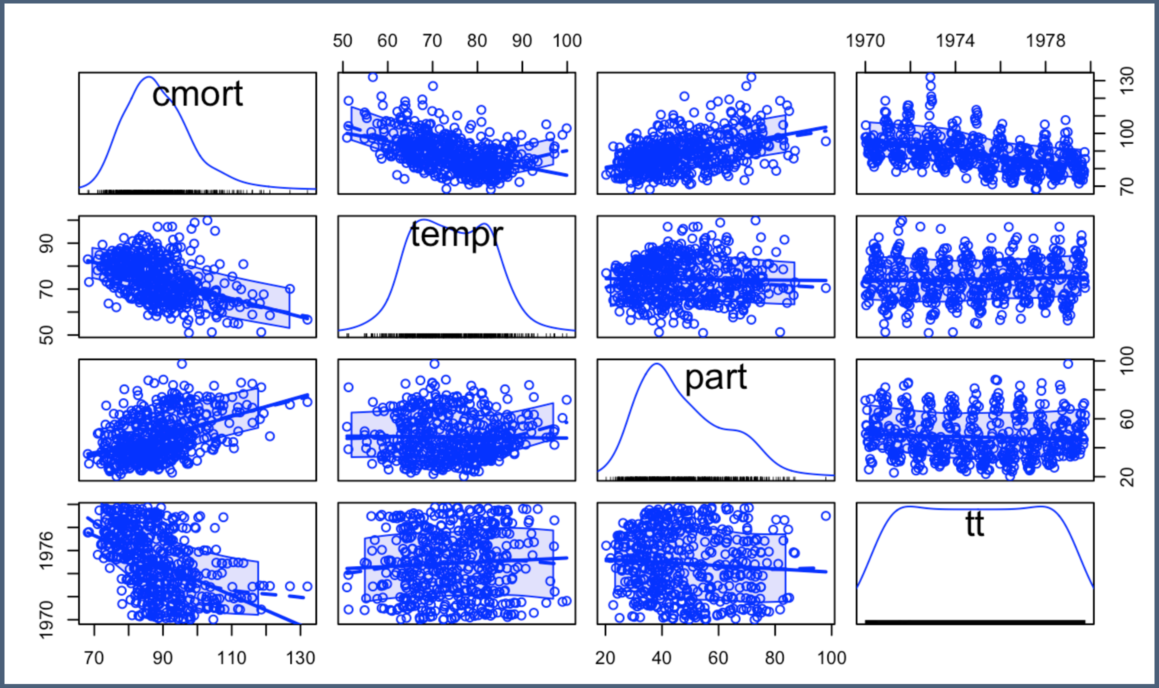
- All pairwise correlations and plots
  - Scatterplot version of a correlation matrix
- Smooth curves
  - Loess curves indicate nonlinear association
  - Not much collinearity (contemporaneous)



Discuss loess estimation of smooth curves shown here in Lecture 9.

Scatterplot Matrix

- One more feature
  - Add the time order so that the last column of plots shows the sequence plots.
  - Already seen these on a prior slide, but useful addition in general.

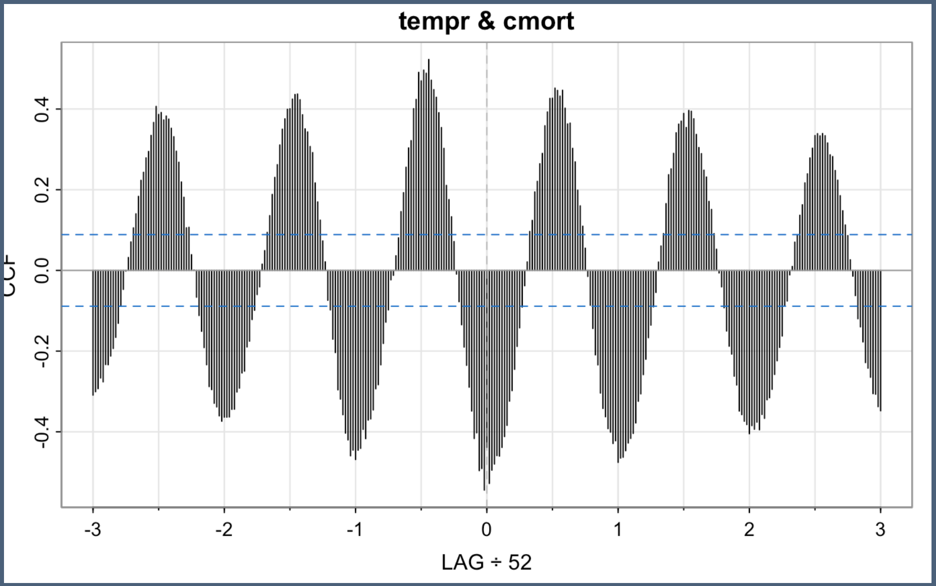


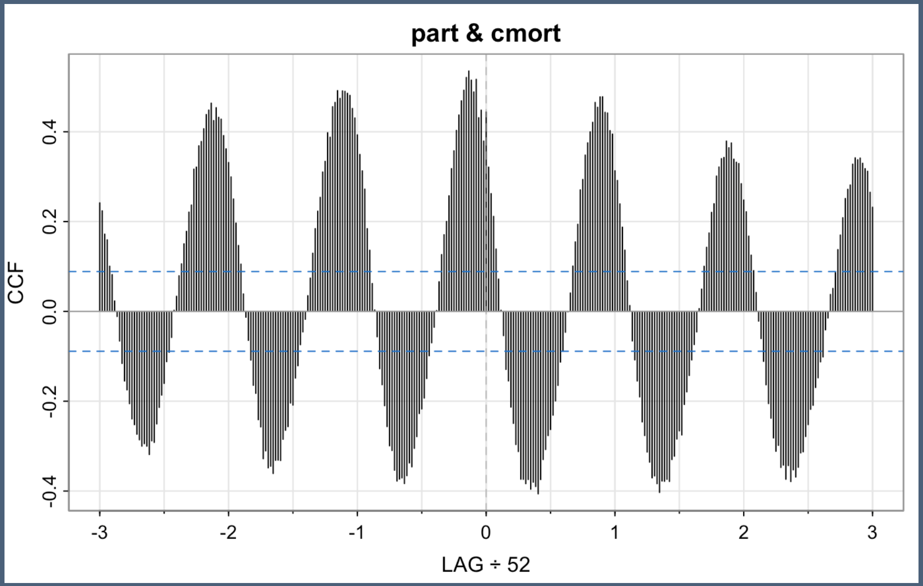
What other features might be relevant?

Check the Lag Structure

Textbook
Example 2.33

- Cross correlations of response and predictors
  - Bivariate only, so may not be the most accurate
  - Appears that need to lag particulates to see best correlation (what happens if you don't?)



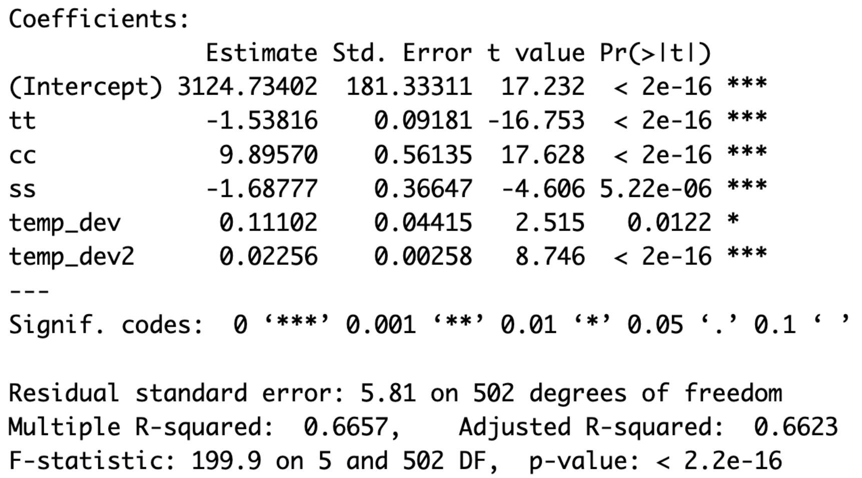


All three variables have annual seasonal variation, and are thus going to be correlated with each other.

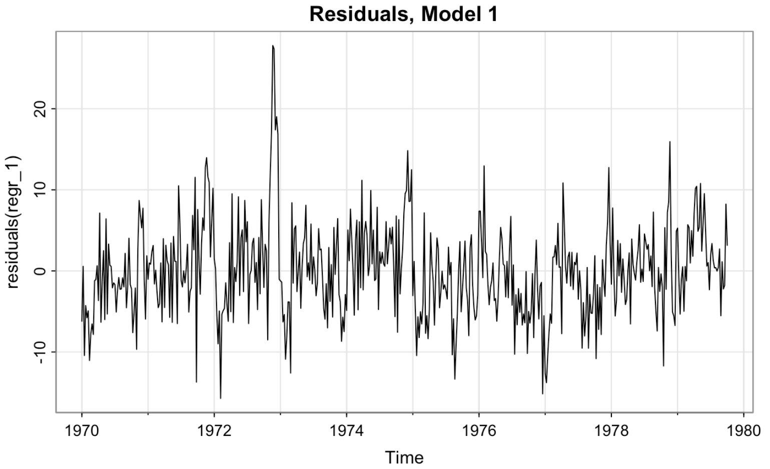
Good habit: List the predictor first so that you'll know which feature is leading and which is lagging

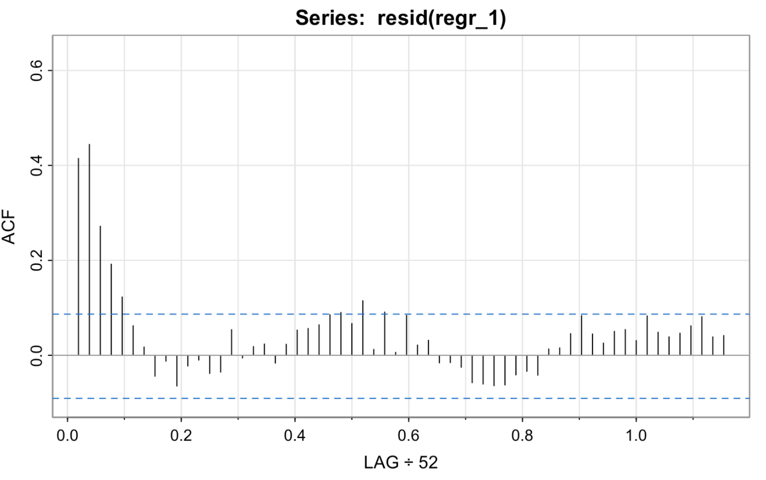
First Model

- Five predictors
  - Trend
  - Annual sinusoid (cosine and sin, text Example 3.15)
  - Temperature deviation from mean
  - Squared deviation
- Diagnostics
  - Substantial annual residual autocorrelation
  - Not surprising that QQ plot is also not great



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 3124.73402  181.33311  17.232  < 2e-16 ***
tt            -1.53816    0.09181 -16.753  < 2e-16 ***
cc             9.89570    0.56135  17.628  < 2e-16 ***
ss            -1.68777    0.36647  -4.606 5.22e-06 ***
temp_dev       0.11102    0.04415   2.515   0.0122 *
temp_dev2      0.02256    0.00258   8.746  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 5.81 on 502 degrees of freedom
Multiple R-squared:  0.6657,    Adjusted R-squared:  0.6623
F-statistic: 199.9 on 5 and 502 DF,  p-value: < 2.2e-16
```

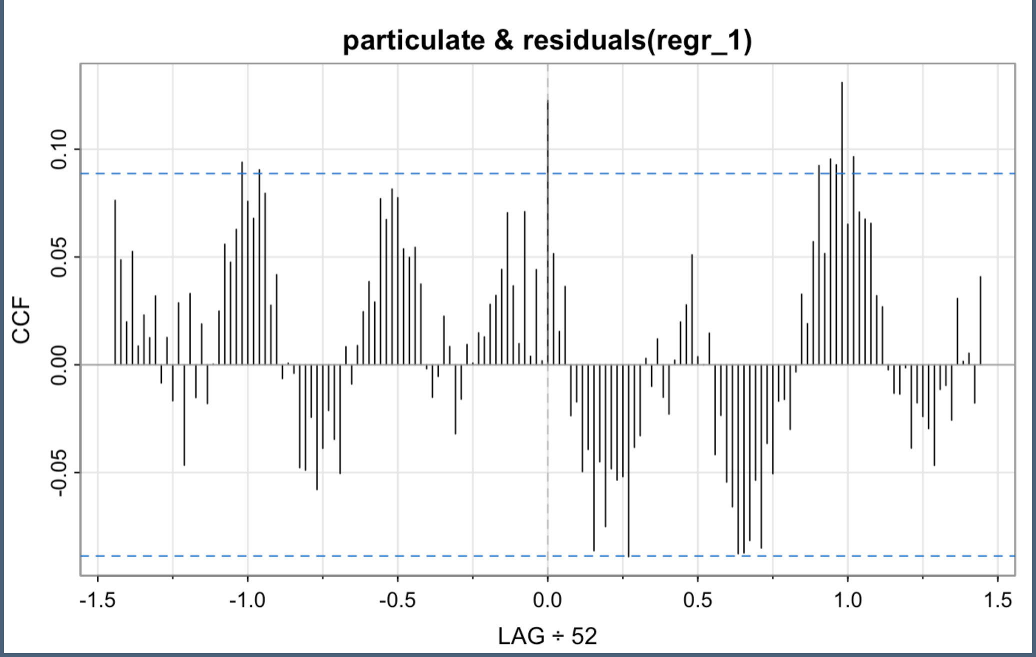




Second Model

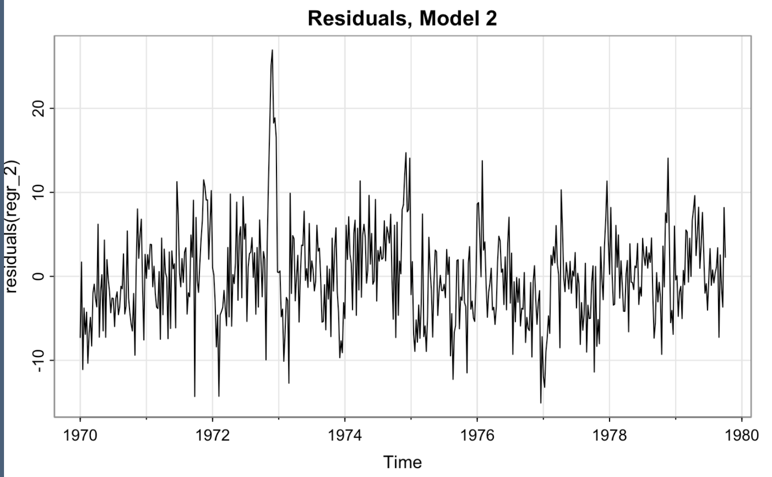
- Six+ predictors
  - Trend and seasonal sinusoid
  - Temperature deviation and its square
  - Particulates, with and without lags

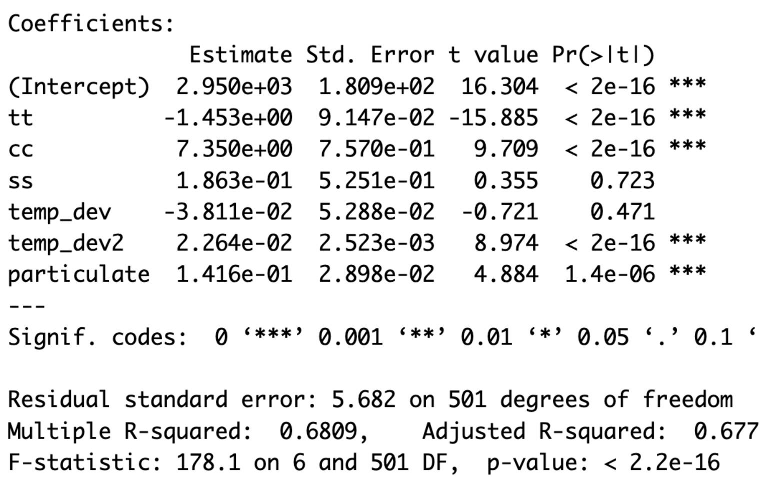
- Which lags of particulates?



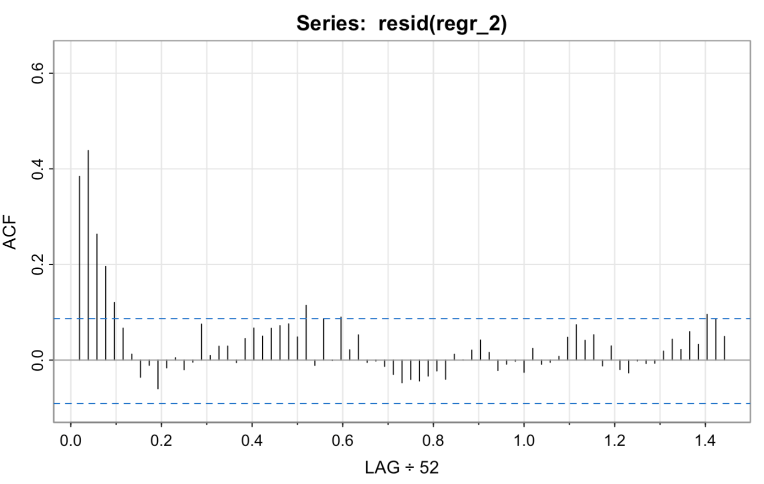
Second Model

- Six predictors
  - Trend, cosine and sine
  - Temperature deviation and its square
  - Particulates at lag 0
- Diagnostics
  - Better fit, but residual autocorrelation remains
  - Inferential statistics are not reliable



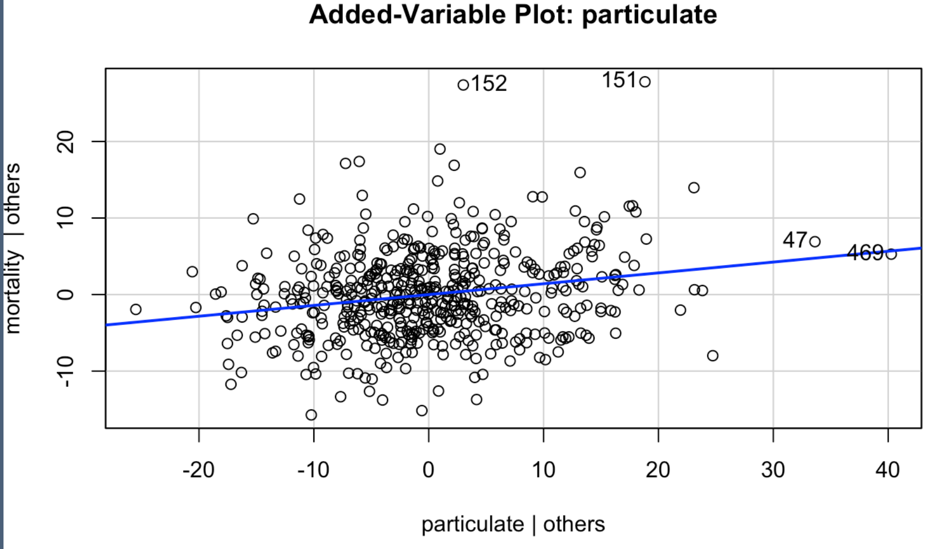


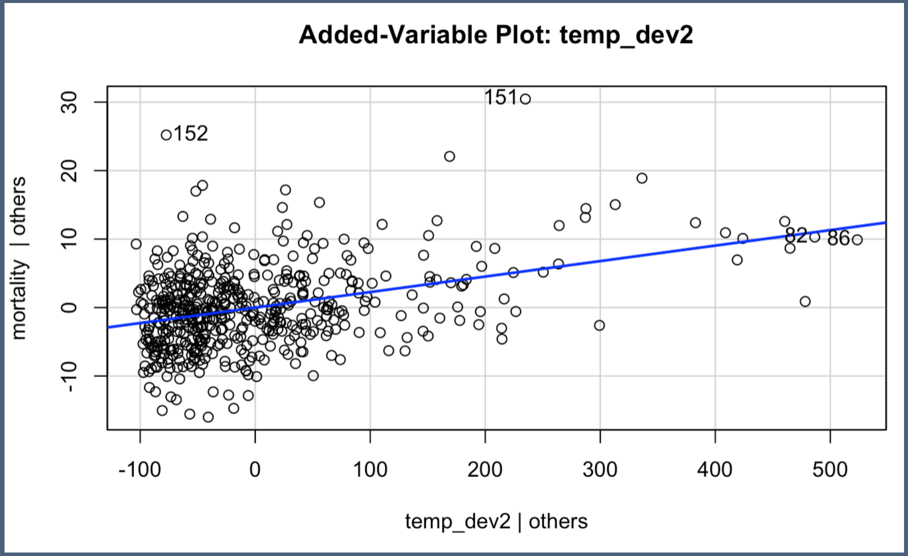
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.950e+03  1.809e+02  16.304  < 2e-16 ***
tt          -1.453e+00  9.147e-02 -15.885  < 2e-16 ***
cc           7.350e+00  7.570e-01   9.709  < 2e-16 ***
ss           1.863e-01  5.251e-01   0.355    0.723
temp_dev    -3.811e-02  5.288e-02  -0.721    0.471
temp_dev2    2.264e-02  2.523e-03   8.974  < 2e-16 ***
particulate  1.416e-01  2.898e-02   4.884  1.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 5.682 on 501 degrees of freedom
Multiple R-squared:  0.6809,    Adjusted R-squared:  0.677
F-statistic: 178.1 on 6 and 501 DF,  p-value: < 2.2e-16
```



Partial Regression Plots

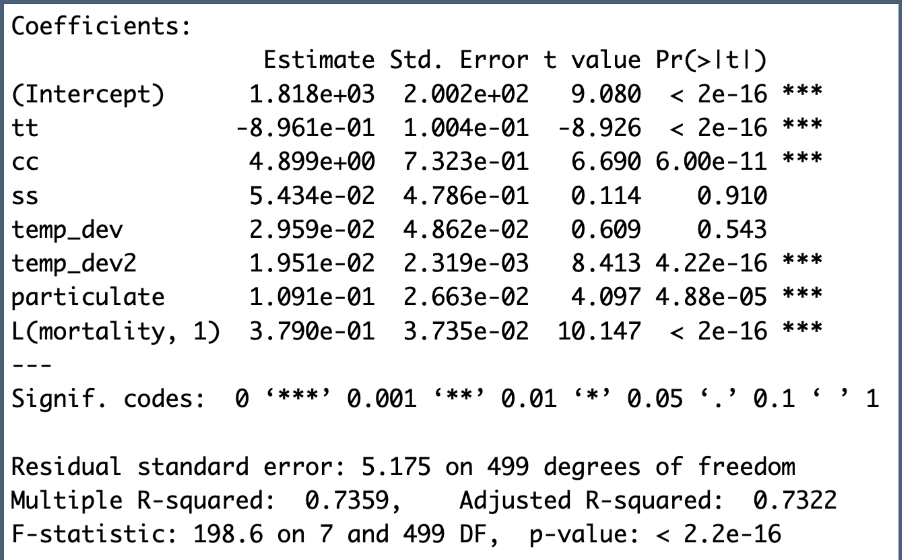
- Particulates and square of temperature
  - Not surprising to see some outliers for the square of the temperature deviation from mean.
  - Nothing particularly unusual in other residual plots (but for the autocorrection)



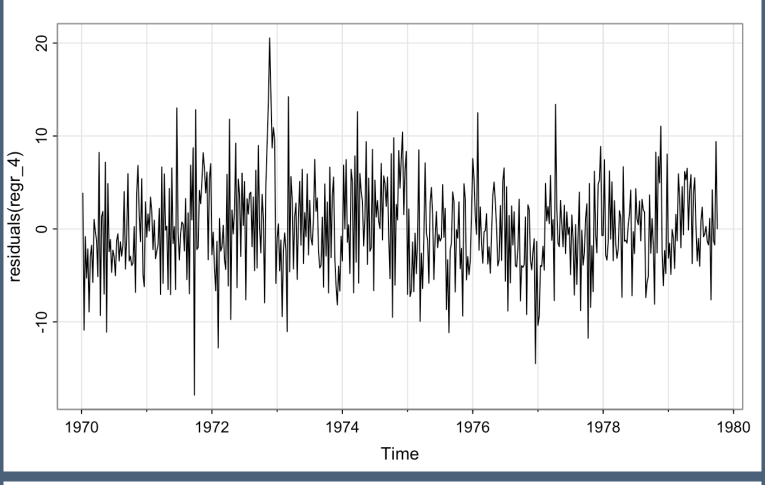


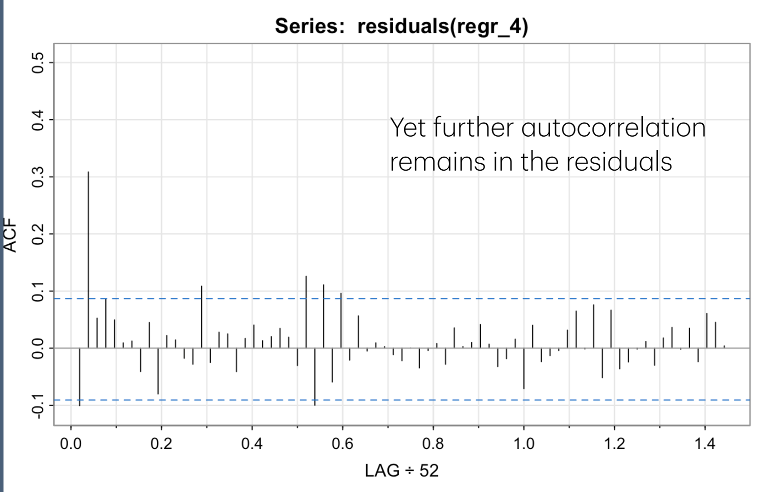
Third Model

- Other lags of predictors offer small gains
  - e.g. Particulates no longer appears a leading indicator
- Lag of the response
  - Lagged endogenous variable improves fit
  - Use the prior mortality itself to predict current mortality
  - Can also motivate differently: use the lag of the residuals



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.818e+03  2.002e+02   9.080  < 2e-16 ***
tt              -8.961e-01  1.004e-01  -8.926  < 2e-16 ***
cc               4.899e+00  7.323e-01   6.690 6.00e-11 ***
ss               5.434e-02  4.786e-01   0.114    0.910    
temp_dev         2.959e-02  4.862e-02   0.609    0.543    
temp_dev2        1.951e-02  2.319e-03   8.413 4.22e-16 ***
particulate      1.091e-01  2.663e-02   4.097 4.88e-05 ***
L(mortality, 1)  3.790e-01  3.735e-02  10.147  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.175 on 499 degrees of freedom
Multiple R-squared:  0.7359,	Adjusted R-squared:  0.7322 
F-statistic: 198.6 on 7 and 499 DF,  p-value: < 2.2e-16
```





Yet further autocorrelation remains in the residuals

Further Models

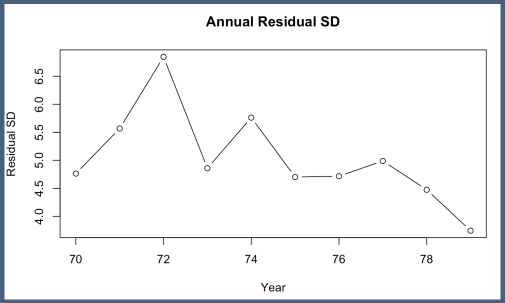
- Example continues
  - Try other lags of predictors (see “Distributed Lag Models” in Maddala)
  - Directly modeling the error process
  - Other types of trends, such as the trend in variability
- Better models
  - Find other substantive variables
  - Manufacture more predictors without gathering more data

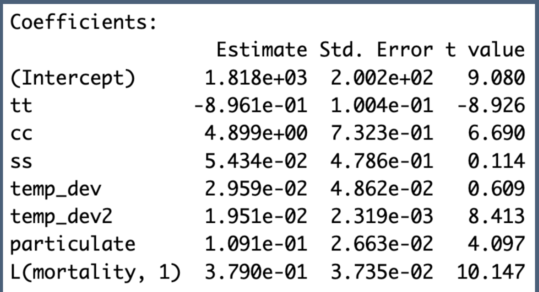
    Nonlinear predictors (e.g. logs or polynomials)

    Interactions: synergy between particulates and temperature

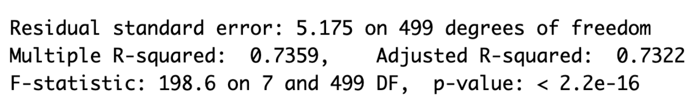
    Additional lags
- Good enough?
  - How well does this model predict claim to predict? $s_w \approx 5.2$

    Is that good enough?
  - What would you need to know to build a forecast?





```
Coefficients:
                 Estimate Std. Error t value
(Intercept)     1.818e+03  2.002e+02   9.080
tt             -8.961e-01  1.004e-01  -8.926
cc              4.899e+00  7.323e-01   6.690
ss              5.434e-02  4.786e-01   0.114
temp_dev        2.959e-02  4.862e-02   0.609
temp_dev2       1.951e-02  2.319e-03   8.413
particulate     1.091e-01  2.663e-02   4.097
L(mortality, 1) 3.790e-01  3.735e-02  10.147
```



```
Residual standard error: 5.175 on 499 degrees of freedom
Multiple R-squared:  0.7359,	Adjusted R-squared:  0.7322
F-statistic: 198.6 on 7 and 499 DF,  p-value: < 2.2e-16
```

What's Next?

- Regression and non-stationary time series

- Converting a non-stationary time series into a stationary series
  - Using a deterministic trend, such as a linear or quadratic time trend
  - Difference the data