Stat 565

Regression With Correlated Errors

Feb 9 2016

Charlotte Wickham

stat565.cwick.co.nz

Projects

Groups of 3-5, I will assign and announce next week.

In weeks 8, 9 & 10 we will have some in class time for working/getting advice on projects.

On form, indicate:

- preference for type of project
- comfort/skill level
- anything else I should know when assigning groups

Proposal due Thursday Feb 25th

Final project due Thursday March 10th

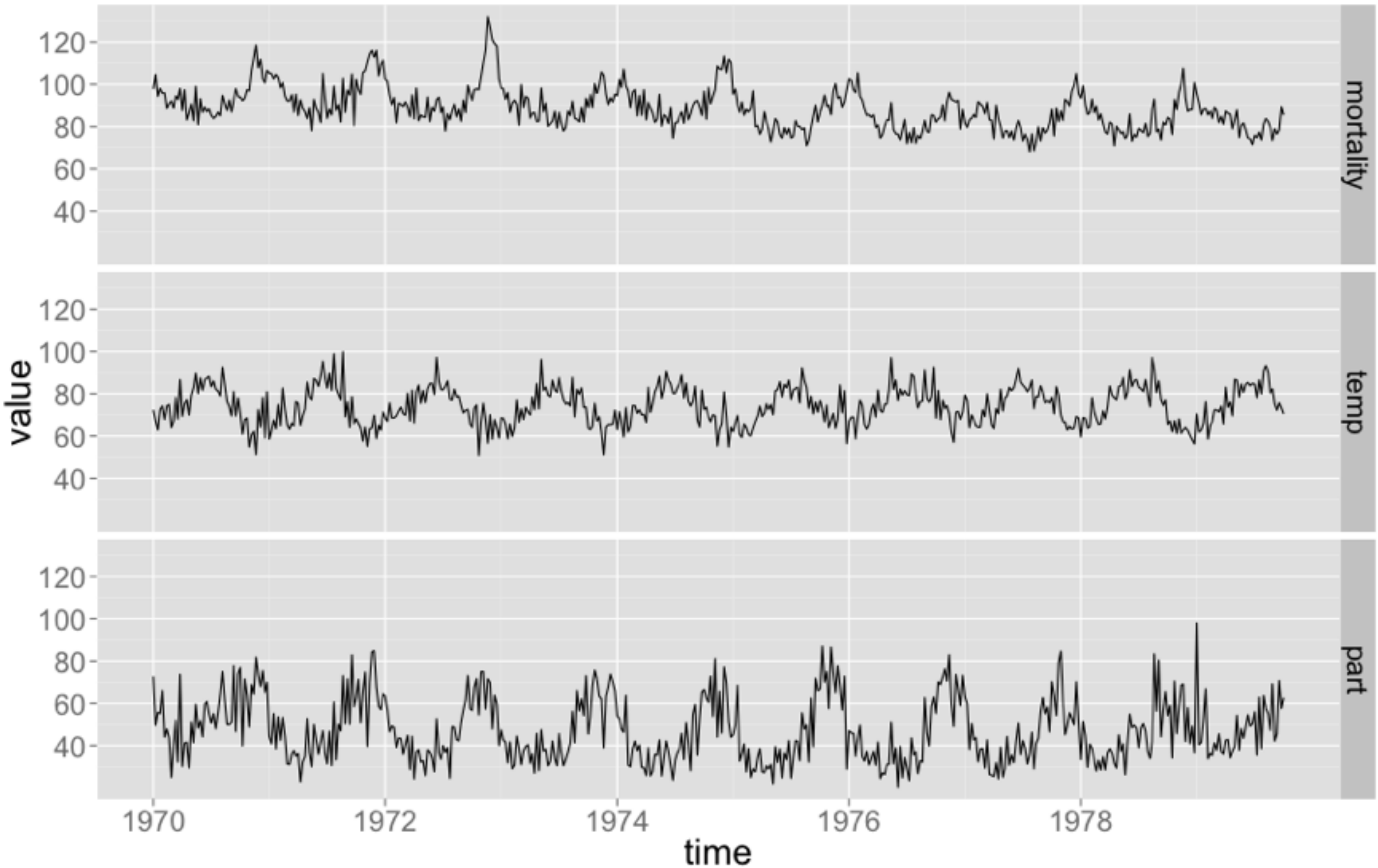
SARIMA models

Good if you are just interested in a short term forecast.

Doesn't result directly in estimates of parameters of the past, i.e. seasonal means, trends, regression parameters...

Today: regression models with correlated errors

Weekly cardiovascular mortality in LA



Can mortality be explained by temperature and particulate matter?

Regression with correlated errors

Use a regression model to explain the non-stationarity in mean.

Extend the usual regression model to allow the errors to be an ARMA process.

Linear Regression Review

A linear regression model, models the response, y_t, as a linear combination of p covariates, x_{t1}, x_{t2}, ..., x_{tp}, and noise, ε_t.

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_p x_{tp} + \varepsilon_t$$

$t = 1, \ldots, n$

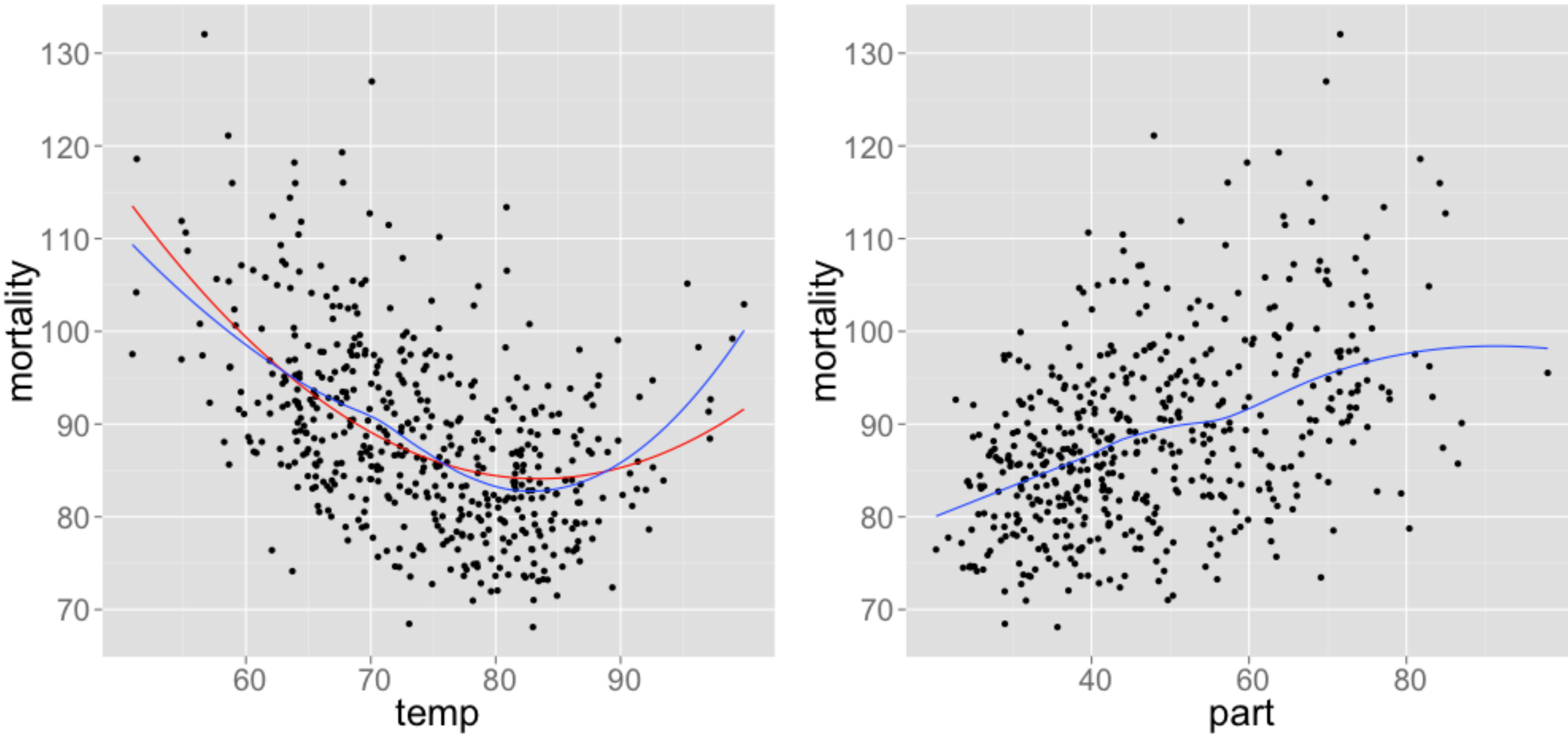
you might be used to seeing i

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

(in matrix notation, y is a nx1 vector of the responses, X is a nxp matrix of covariates, β a px1 vector of parameters, ε is a nx1 vector of errors)

Your turn

What are the assumptions of ordinary least squares regression?



Suggests the regression model:

$$\text{mortality}_t = \beta_0 + \beta_1 t + \beta_2 \text{temp}_t + \beta_3 \text{temp}_t^2 + \beta_4 \text{part}_t + \varepsilon_t$$

```
> fit_lm <- lm(mortality ~ time0 + temp_sc + temp_2 + part, data = mort)
```

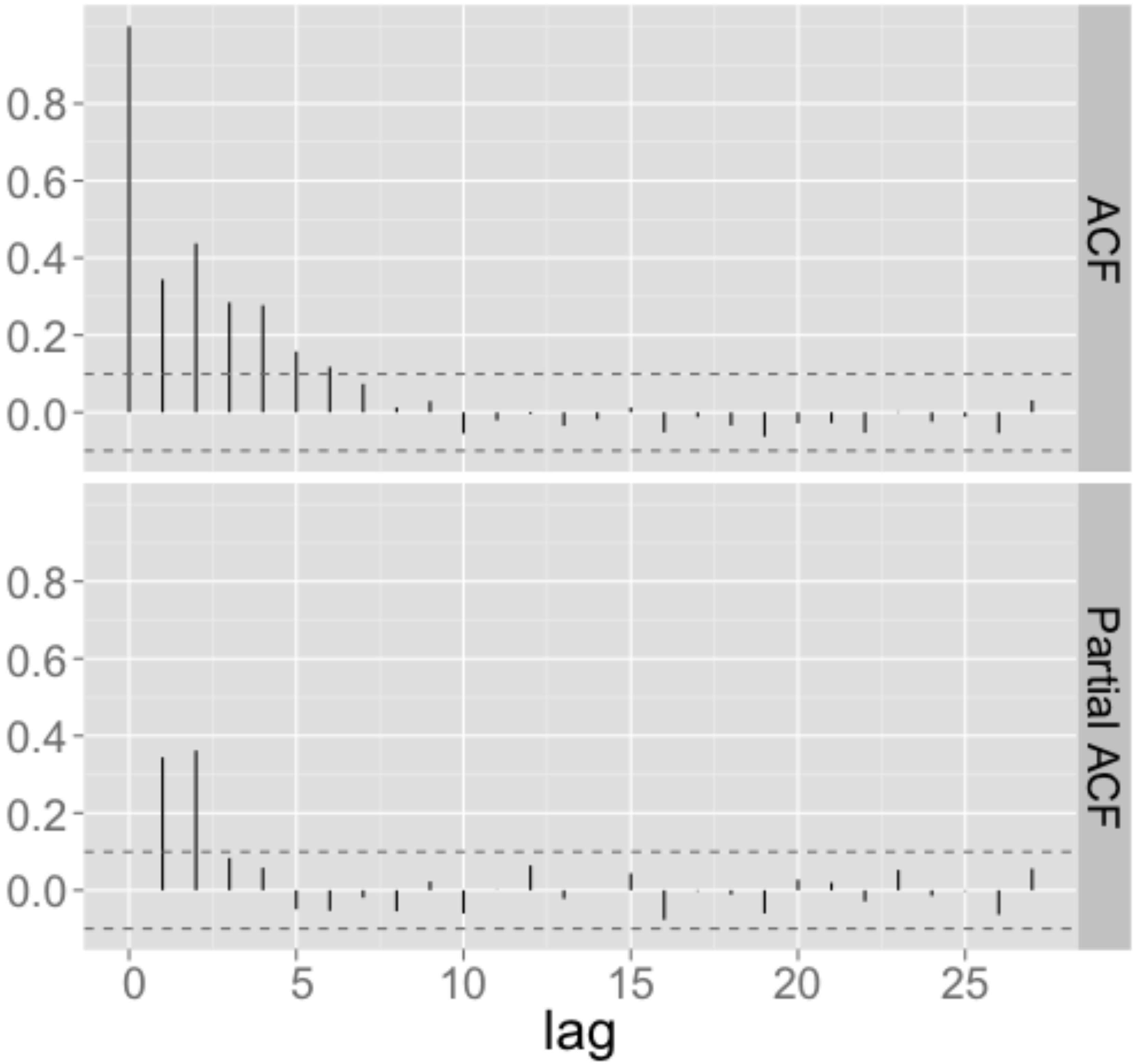
I started time at 0, centered temp about it's mean
and found the square of temp

```
> summary(fit_lm)

Call:
lm(formula = mortality ~ time0 + temp_sc + temp_2 + part, data = mort)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.0760  -4.2153  -0.4878   3.7435  29.2448 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 81.565394   1.101148   74.07  < 2e-16 ***
time0       -1.395901   0.101009  -13.82  < 2e-16 ***
temp_sc     -0.472469   0.031622  -14.94  < 2e-16 ***
temp_2       0.022588   0.002827    7.99 9.26e-15 ***
part         0.255350   0.018857   13.54  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.385 on 503 degrees of freedom
Multiple R-squared: 0.5954,	Adjusted R-squared: 0.5922 
F-statistic:   185 on 4 and 503 DF,  p-value: < 2.2e-16
```

Assumptions

Everything looks good, except....

```
examine_corr(residuals(fit_lm))
```



Assumptions say this should be white noise, looks like ?

Generalized Least Squares

Instead of variance $\sigma^2 I$, allow errors to have covariance matrix Σ.

If Σ is known multiplying by $\Sigma^{-1/2}$, reduces the problem to the usual regression problem.

If Σ is unknown, start with estimating the β using OLS, use the residuals to estimate Σ and iterate. Should converge to the MLEs (under Normal errors).

Correlated errors model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_p x_{tp} + z_t$$

z_t is a stationary ARMA(p,q) process

Or equivalently $z_t \sim N(0, \Sigma)$ where,

$\Sigma_{ij} = Cov(z_i, z_j) = \Upsilon(| i - j |)) =$ some function of σ^2, β and α

Plus usual assumptions:

linearity, z_t independent of X.

ARMA errors

If the errors are an ARMA process, Σ, can be written in terms of our parameters β and α.

(Since $\Sigma_{ij} = \mathrm{Cov}(z_i, z_j) = \gamma(|i - j|)$)

If we specify p and q, Σ is known up to β and α.

and can be estimated by GLS.

The standard errors on the regression coefficients, β, depend on X, σ^2, β and α.

Procedure

1. Fit the model for the mean using OLS.
2. Examine the residuals to identify an appropriate ARMA(p, q) process.
3. Fit the GLS model.
4. Model diagnostics.
5. Interpret (forecast?).

In R

Either (a little more general, i.e. not just for time series):

```
library(nlme)
gls(y ~ x1 + x2,
  correlation = corARMA(p = p, q = q),
  method = "ML")
```

Or (faster and can handle seasonal arima models):

```
arima(y, order = c(p, d, q), xreg = X)
```

We already identified the errors in the mortality series as AR(2).

```
gls_fit <- gls(mortality ~ time0 + temp_sc + temp_2 + part,
              data = mort,
              correlation = corARMA(p = 2), method = "ML")
```

or

```
arima_fit <- with(mort, arima(mortality, order = c(2, 0, 0),
                 xreg = cbind(time0, temp_sc, temp_2, part)))
```

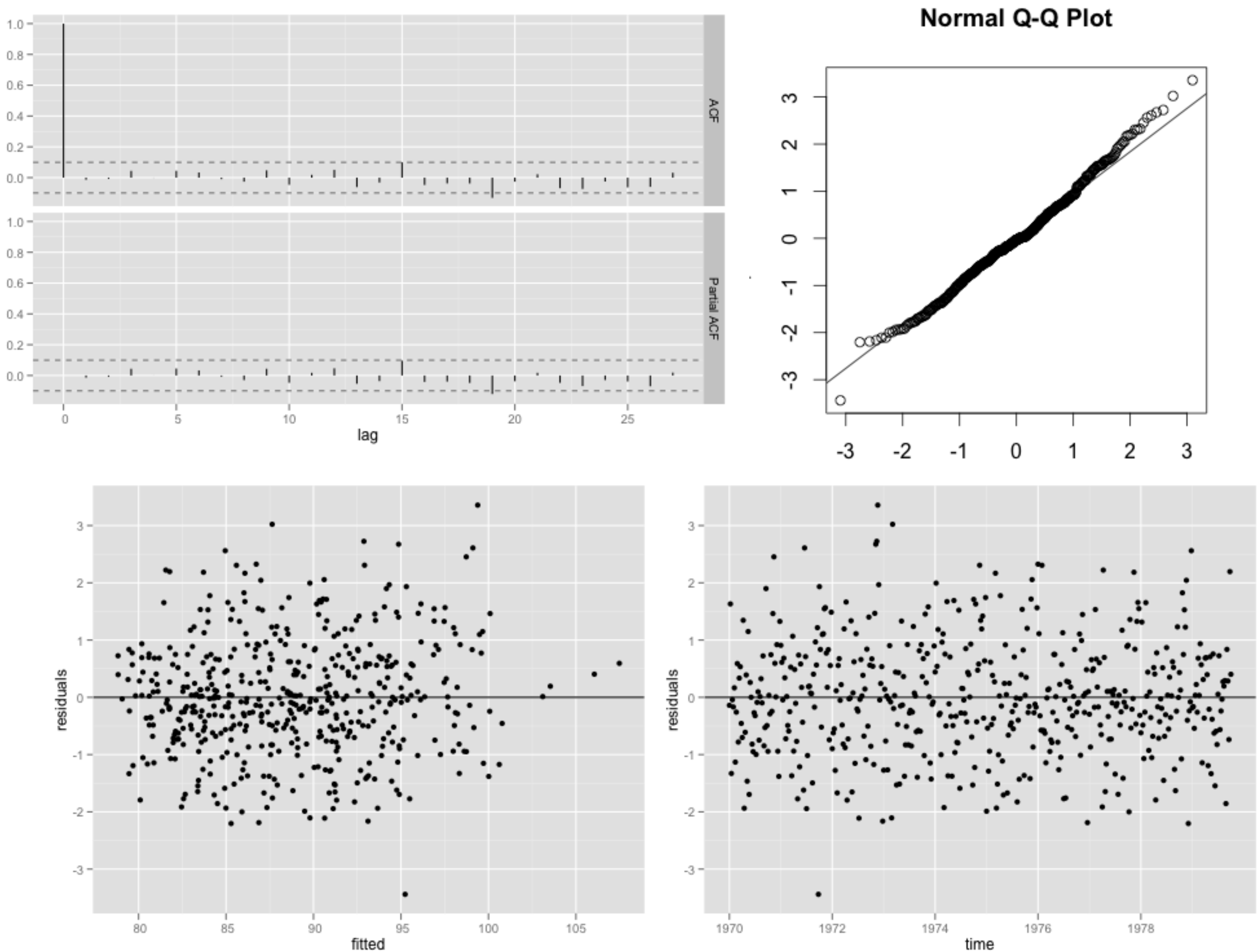
AR(2): $z_t = \alpha_1 z_{t-1} + \alpha_2 z_{t-2} + w_t$

Diagnostics

`residuals(arima_fit)` gives estimates of w_t, the white noise.

`residuals(gls_fit)` gives estimates of z_t, the ARMA process.

use `residuals(gls_fit, type = ”normalized”)` to get w_t


ACF
Partial ACF
lag
Normal Q-Q Plot
residuals
fitted
residuals
time

wrong

Ordinary linear regression

Assume white noise errors

```
> round(confint(fit_lm), 2)
            2.5 % 97.5 %
(Intercept) 79.40  83.73
time0       -1.59  -1.20
temp_sc     -0.53  -0.41
temp_2       0.02   0.03
part         0.22   0.29
```

right

Linear regression with correlated errors

Assume AR(2) errors

```
> round(intervals(gls_fit)$coef, 2)
            lower  est. upper
(Intercept) 82.25 87.61 92.96
time0       -2.35 -1.52 -0.68
temp_sc     -0.10 -0.02  0.07
temp_2       0.01  0.02  0.02
part         0.11  0.15  0.20
```

```
> round(intervals(gls_fit)$corStruct, 2)
     lower est. upper
Phi1  0.35 0.38  0.38
Phi2  0.35 0.43  0.51
```

How would you interpret the coefficient on time0?

Your turn

Where does most of the variation in temperature come from?

What about particulates?

What about mortality?