AS.280.347
CLASS 2.3

- Review your work!
- Pooling results from different cities
- Your projects!

Assignment 2.2

1. Update your time series display of PM10, temperature, and total mortality versus date.

2. For each city, regress mortality on PM10 using a log-linear (Poisson) model with different indicator variables for time:
   - A: death ~ pm10 0 degrees of freedom (df)
   - B: death ~ pm10 + as.factor(season) 4-1 = 3 df
   - C: death ~ pm10 + as.factor(month) 12-1 = 11 df
   - D: death ~ pm10 + as.factor(season)*as.factor(year) 4*19-1 = 75 df
   - E: death ~ pm10 + as.factor(month)*as.factor(year) 12*19-1 = 227 df

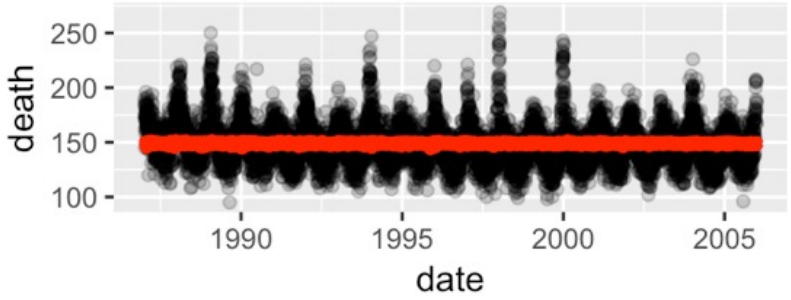
3. Plot mortality against time and add a continuous line for each predicted model.

4. Display the five PM10 coefficients (A-E) with confidence intervals in a table or graph to see the effect of the method of control for seasonality.

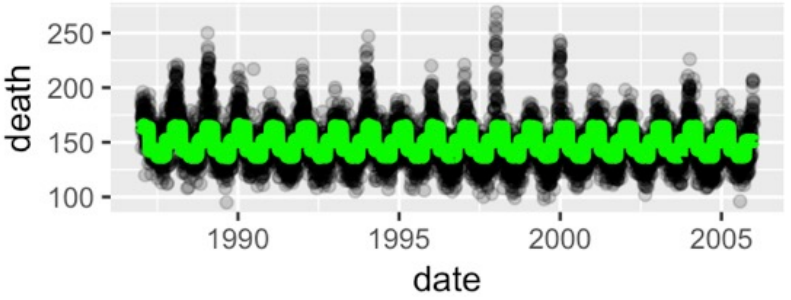
5. Repeat the regression models (A-E) with natural splines to give a smooth relationship between mortality and time using ns(time, df) for df = 0, 3, 11, 75, 227. Plot the data and predicted curves once again.

LA plots fitted to different models

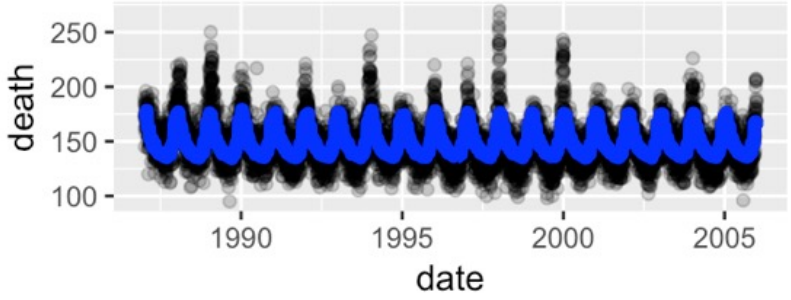
PM10



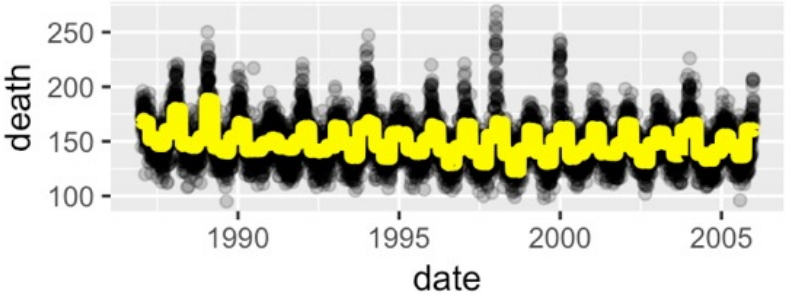
pm10 + as.factor(season)



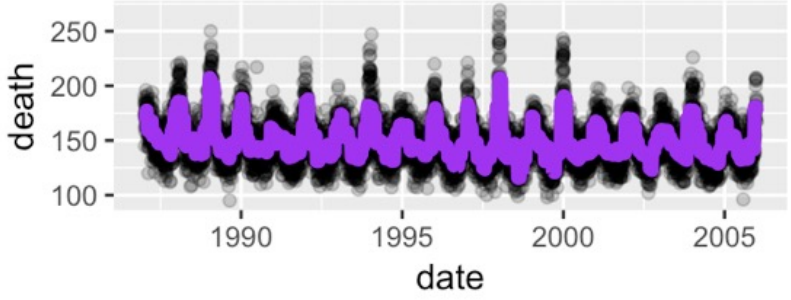
pm10 + as.factor(month)



pm10 + as.factor(season)*as.factor(



pm10 + as.factor(month)*as.factor(year)



```
la.data$predA = exp(pred.modelA)
la.data$predB = exp(pred.modelB)
la.data$predC = exp(pred.modelC)
la.data$predD = exp(pred.modelD)
la.data$predE = exp(pred.modelE)


modelA<-ggplot(aes(date, death), data=la.data) +
  geom_point(alpha=0.2)+ ggtitle("PM10") +
  geom_line(data=la.data[!is.na(la.data$predA),],aes(x=date, y=predA),
                                                          colour="red", size=2)


modelB<-ggplot(aes(date, death), data=la.data) + ggtitle("pm10 + as.factor(season)") +
  geom_point(alpha=0.2) +
  geom_line(data=la.data[!is.na(la.data$predB),],aes(x=date, y=predB), colour="green", size=2)


modelC<-ggplot(aes(date, death), data=la.data) + ggtitle("pm10 + as.factor(month)") +
  geom_point(alpha=0.2) +
  geom_line(data=la.data[!is.na(la.data$predC),],aes(x=date, y=predC), colour="blue", size=2)


modelD<-ggplot(aes(date, death), data=la.data) +
     ggtitle("pm10 + as.factor(season)*as.factor(year)") +
     geom_point(alpha=0.2) + geom_line(data=la.data[!is.na(la.data$predD),],
                                       aes(x=date, y=predD), colour="yellow", size=2)


modelE<-ggplot(aes(date, death), data=la.data) + ggtitle("pm10 + as.factor(month)*as.factor(year)") +
     geom_point(alpha=0.2)+
     geom_line(data=la.data[!is.na(la.data$predE),],aes(x=date, y=predE), colour="purple", size=2)


grid.arrange(modelA, modelB, modelC, modelD, modelE , nrow=3)
```

Table 1: This table displays the results of the poisson regression analysis according to each of the five models. 95% confidence intervals contain a lower bound below 0 for most models, with only models C and D consistently displaying intervals entirely above 0 and p values less than 0.05. Even in these cases, however, the Beta 1 coefficients are miniscule, suggesting a small effect overall.

	Log RMR (pm10)	p	95% CI
Mortality ~ pm10 (NY)	-0.00002	0.90870	(-0.00038, 0.00034)
Mortality ~ pm10 + season (NY)	0.00119	0.00000	(-0.00061, -0.00008)
Mortality ~ pm10 + month (NY)	0.00123	0.00000	(0.00051, 0.00123)
Mortality ~ pm10 + season x year (NY)	0.00133	0.00000	(0.00082, 0.00155)
Mortality ~ pm10 + month x year (NY)	0.00124	0.00000	(-0.00010, 0.00044)
Mortality ~ pm10 (LA)	-0.00035	0.01004	(-0.00023, 0.00052)
Mortality ~ pm10 + season (LA)	0.00017	0.22028	(0.00085, 0.00161)
Mortality ~ pm10 + month (LA)	0.00032	0.02268	(0.00004, 0.00059)
Mortality ~ pm10 + season x year (LA)	0.00020	0.16701	(-0.00028, 0.00049)
Mortality ~ pm10 + month x year (LA)	0.00038	0.01209	(0.00095, 0.00170)
Mortality ~ pm10 (SEA)	0.00087	0.00000	(-0.00008, 0.00047)
Mortality ~ pm10 + season (SEA)	0.00015	0.43603	(-0.00025, 0.00053)
Mortality ~ pm10 + month (SEA)	0.00011	0.58705	(0.00082, 0.00166)
Mortality ~ pm10 + season x year (SEA)	0.00014	0.46830	(0.00008, 0.00068)
Mortality ~ pm10 + month x year (SEA)	0.00015	0.48014	(-0.00026, 0.00055)

Module 2: Particulate air pollution and mortality

- Question 2.1 (Q2.1): How does the daily risk of death depend upon air pollution level in American cities?
- Question 2.2 (Q2.2): Is the estimate of the pollution effect sensitive to assumptions about seasonal or weather effects?
- **Question 2.3 (Q2.3): How do you pool PM effect (log relative rate) estimates from multiple cities taking account of both natural geographic variability in the true effects and statistical errors that might differ among cities?**
- We will answer these questions using data from the National Morbidity and Mortality Air Pollution Study (NMMAPS)

Pooling the city-specific estimates

- Recall that Y counts the number of events in a fixed time period (for us, number of deaths in one day)
- Let μ = mean(Y) = mean "rate" of events per day in the time period (for us, mean daily mortality)
- We can often model Y as a Poisson distribution with mean μ
- Model equation:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Pooling the city-specific estimates

- For each city, we fit a model and estimate the log relative rate

$$\hat{\beta}_{Chicago} \quad \hat{\beta}_{NY} \quad \hat{\beta}_{SLC}$$

- How do we interpret this log relative rate?

$$\hat{\beta}_{NY} = \log\left(\frac{daily\ mortality\ in\ NY for\ PM10 = x + 1}{daily\ mortality\ in\ NY\ for\ PM10 = x}\right)$$

$$= \log\left(\frac{(daily\ mortality\ in\ NY for\ PM10 = x+1)/PopulationNY}{(daily\ mortality\ in\ NY\ for\ PM10 = x)/PopulationNY}\right)$$

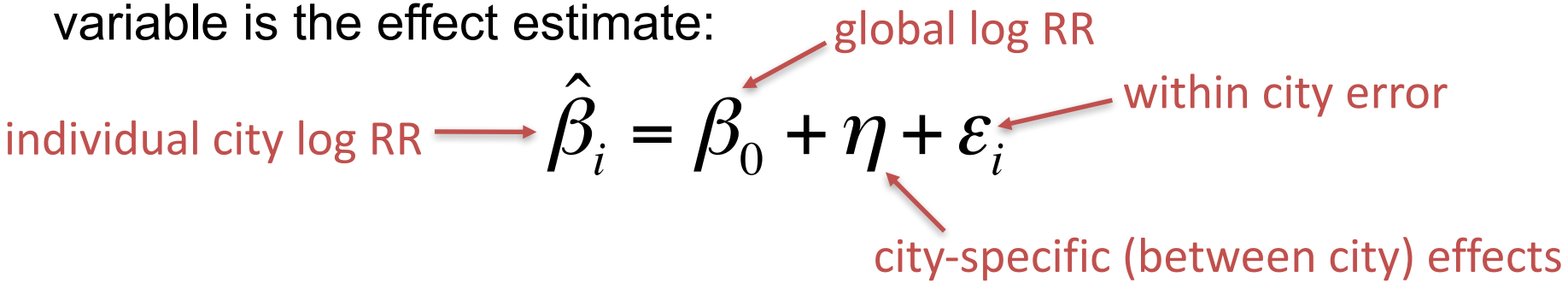
$$= \log\left(\frac{daily\ mortality\ \boldsymbol{rate}\ in\ NY for\ PM10 = x+1}{daily\ mortality\ \boldsymbol{rate}\ in\ NY\ for\ PM10 = x}\right)$$

Pooling the city-specific estimates

- For each city, we fit a model and estimate the log relative rate

$$\hat{\beta}_{Chicago} \quad \hat{\beta}_{NY} \quad \hat{\beta}_{SLC}$$

- We are interested in an estimate of a "global" log relative rate across all cities, but we realize that there are some "city-specific" characteristics that means the rate will vary from city to city!
- We can do this using "meta-regression", where the outcome variable is the effect estimate:



- We still need to weight each estimate by it's precision (inverse variance)!

Pooling the City-Specific Estimates

City	Estimate of Log Rel Rate	$SE_{Estimate}$	Total Variance ($V=t^2+SE^2$)	V^{-1}	$V^{-1}/Sum(V^{-1})$	Weighted Average

$$\hat{\beta}_i = \beta_0 + \eta + \varepsilon_i$$

$$Var(\hat{\beta}_i) = Var(\beta_0 + \eta + \varepsilon_i) = Var(\eta) + Var(\varepsilon_i) = \tau^2 + SE^2$$

$$\hat{\tau}^2 = Var(\hat{\beta}_i) - Average(SE^2)$$

Assignment 2.3

- **For each single city:**
  - Update your time series display of PM10, temperature, and total mortality versus date.
  - Regress mortality on PM10 using a Poisson model with different indicator variables for time: (a) nothing (0 df); (b) season (4-1 df); (c) month (12-1 df); (d) season by year (4x19-1 df); (e) month by year (12x19-1 df)
  - Repeat the regressions with natural splines to give a smooth relationship between mortality and time using ns(time, df) for df = 0, 3, 11, 75, 227
  - Display the 10 PM10 coefficients with confidence intervals in a table or graph to see the effect of the method of control for seasonality.
  - Choose your preferred model from among the 10 above. Explain your choice. **Interpret the log relative rate from this chosen model.**
- **Across the cities:**
  - Using your chosen seasonal adjustment – the same for each city – pool the city-specific estimates of log relative rate from your team to estimate an average value taking account of statistical and city-specific variation in the log relative rate estimates.
  - Write an R function to do the pooling calculation

- Work together in groups!
- Submit assignment in R markdown through Blackboard by **Sunday, March 24th** @ midnight.

Thinking ahead: your project!

- Question:
- Data set and design
  - Outcome:
  - Predictor variables of primary interest:
  - Effect modifiers:
  - Confounders:
- Directed Acyclic Graph (DAG):
- Primary analysis to address question:
- Communicating results in tables and figures:

Thinking ahead: your project!

By the end of Spring Break, think about:

- A research question of interest in public health
- A data source that you can use to answer this question

Framing a research question in public health:

- Start with a general area of public health in which you have interest, and then narrow to a specific question you'd like to answer.
- It can be helpful to frame your question in terms of investigating a relationship between a specific outcome variable (like "disease status" for our Module 1) and one or more primary predictor variables ("smoking status" for our Module 1.)
- Later you will need to think about the possibility of effect modifiers and possible confounders, but for now just think about that primary relationship of interest!

Locating data to answer this question:

- If you have a specific area of interest in mind, you can Google for data in that area
- Or explore the links below to see what type of data is available:

  https://www.healthdata.gov/
  http://guides.lib.berkeley.edu/publichealth/healthstatistics/rawdata
  http://www.datasciencecentral.com/profiles/blogs/10-great-healthcare-data-sets
  https://www.cdc.gov/nchs/data_access/ftp_data.htm
  https://catalog.data.gov/dataset?_organization_limit=0&organization=hhs-gov#topic=health_navigation