Poisson Regression

Patrick Breheny

April 19

Count data

- Count data is another common type of data in observational and epidemiological studies
- This type of data naturally arises from studies investigating the incidence or mortality of diseases in a population
- The Poisson distribution is a natural choice to model the distribution of such data
- As we will see, it is also sometimes convenient to model cohort studies using the Poisson distribution

Poisson regression

- As with the binomial distribution leading to logistic regression, a simple Poisson model is quite limited
- We want to allow each sampling unit (person, county, etc.) to have a unique rate parameter λ_i, depending on the explanatory variables
- The random and systematic components are as follows:
  - Random component: $y_i \sim \text{Pois}(\lambda_i)$
  - Systematic component: $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$

Poisson regression: Link function

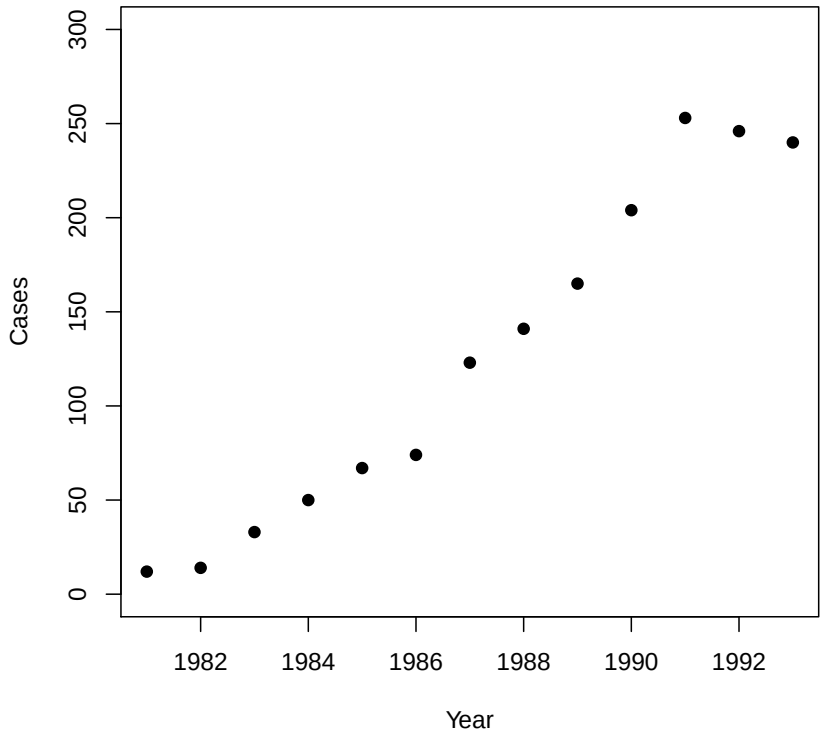
- Recall that the canonical link for the Poisson distribution is the log link
- Thus,

$$\log(\lambda_i) = \eta_i$$
$$\lambda_i = \exp(\eta_i)$$

- Note again that the canonical link ensures that $\lambda_i > 0$, as it must be for the Poisson distribution

Belgian AIDS data

As a first example of Poisson regression, consider the following data on the number of new cases of AIDS in Belgium, 1981–1993:



Modeling the Belgian AIDS data

- Consider the following simple model:

$$\eta_i = \beta_0 + \beta_1 \texttt{Year}$$

- As we have remarked previously, this is equivalent to fitting the exponential growth model

$$\lambda_i = \gamma \exp(\delta t_i),$$

  where $\beta_0 = \log(\gamma)$ and $\beta_1 = \delta$
- Exponential growth models are reasonable in the early stages of an epidemic

Model fitting and inference

- Fitting these models (as you know from the homework) can be accomplished via an iteratively reweighted least squares algorithm, with the reweighting step

$$w_i^{(m)} = \hat{\lambda}_i^{(m)}$$

- Furthermore (as you also know from the homework), we can carry out inference according to the Wald approximation

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\right)$$

- We can then transform estimates and confidence intervals to get inference on the λ scale, just as we did for logistic regression

Poisson regression in SAS/R

- Fitting these models in SAS and R is straightforward
- In SAS,

```
PROC GENMOD DATA=aids;
  MODEL Cases = Year / DIST=POI;
RUN;
```

- In R

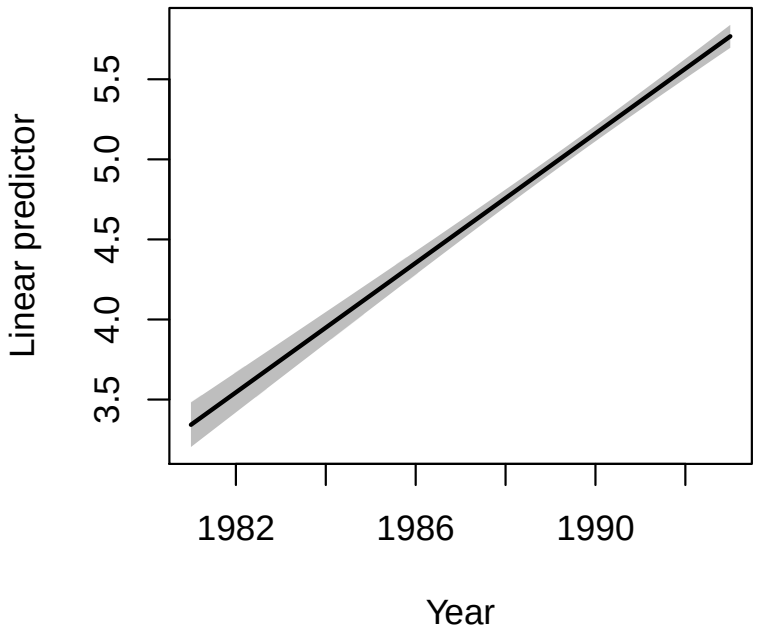
```
glm(Cases~Year,aids,family=poisson)
```

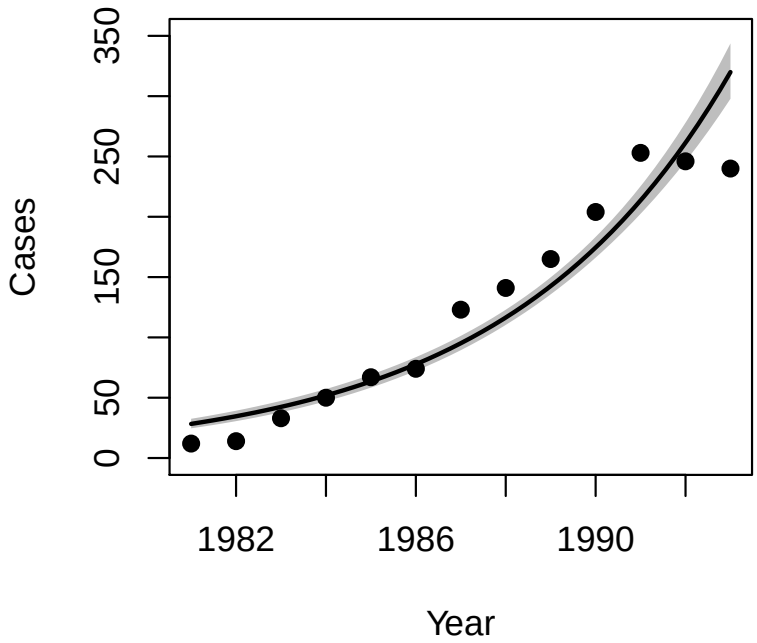
Likelihood ratio intervals and tests

- Again, the default output is Wald-style inference
- To obtain likelihood ratio tests and confidence intervals in SAS, one can add the options `LRCI` and `TYPE3` to the `MODEL` statement
- In `R`, the `confint` function again produces likelihood ratio intervals, while likelihood ratio tests can be carried out by fitting the full model (`fit`) and the reduced model (`fit0`), then submitting

```
anova(fit0,fit,test="Chisq")
```

Results


Linear predictor
3.5
4.0
4.5
5.0
5.5
1982
1986
1990
Year


Cases
0
50
150
250
350
1982
1986
1990
Year

Pearson residuals

- As with logistic regression, there are two commonly used types of residuals for Poisson regression: Pearson residuals and deviance residuals
- Pearson residuals are straightforward:

$$r_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

- Note that if we call y_i the observed quantity and $\hat{\lambda}_i$ the expected quantity, we have

$$\sum_i r_i^2 = \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}},$$

the usual χ^2 test statistic

Deviance

- Before we derive the deviance residuals, we need to revise the informal, oversimplified definition of deviance that I provided earlier
- *Deviance* is defined as twice the difference in log-likelihood between a model and an optimal model for which $\hat{\mu}_i = y_i$ for all observations; denoting these quantities ℓ and $\ell_{\max}$:

$$D = 2(\ell_{\max} - \ell)$$

- This detail was not relevant to our earlier uses of deviance, as for the Bernoulli and normal distributions, $\ell_{\max} = 0$
- This is not the case for the Poisson distribution, however

Deviance residuals

- For the Poisson distribution,

$$d_i = s_i \sqrt{2\{y_i \log(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)\}},$$

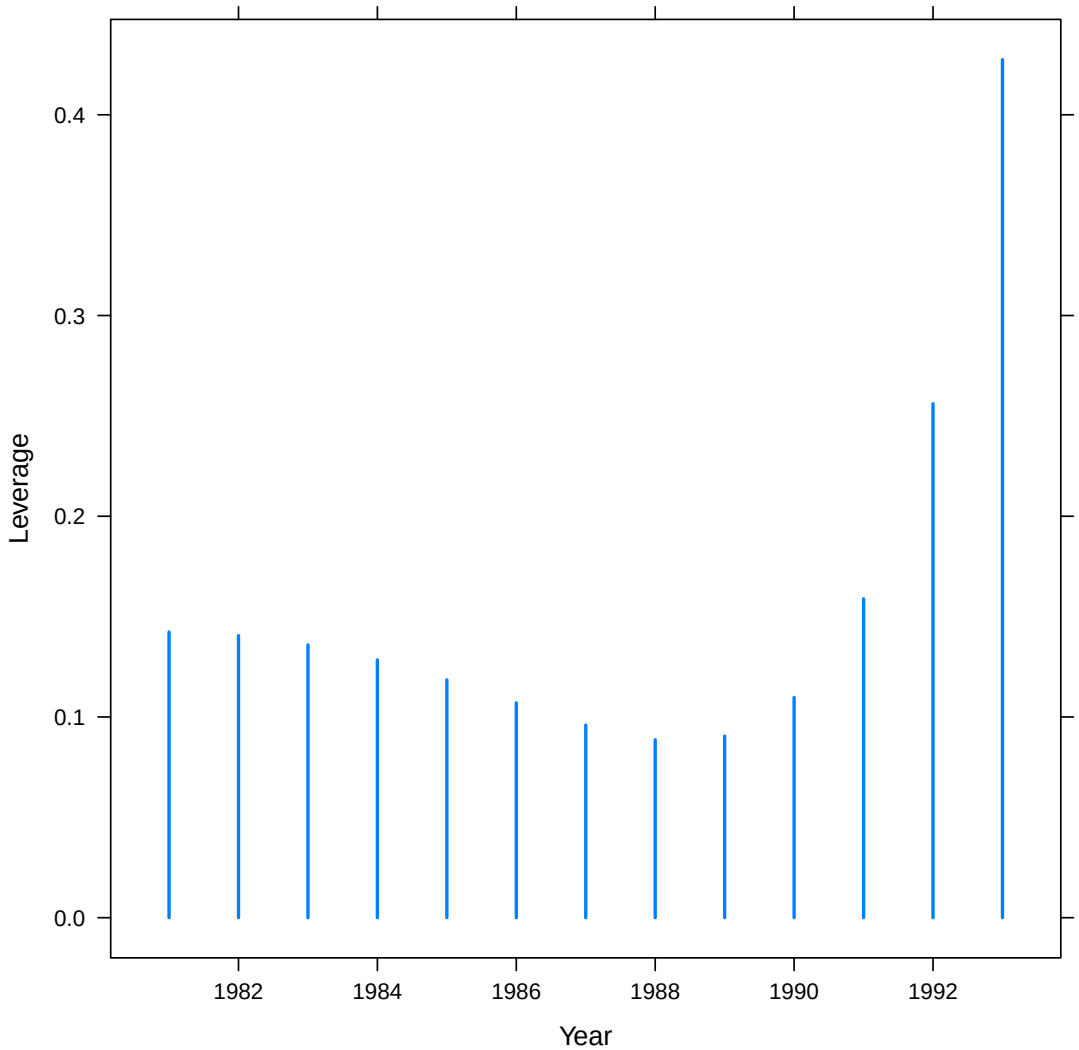
  where you may recall that s_i was the sign of $y_i - \hat{\lambda}_i$
- The deviance is $D = \sum_i d_i^2$, although if the model has an intercept, then $\sum_i y_i = \sum_i \hat{\lambda}_i$, and the deviance simplifies to

$$D = 2 \sum_i y_i \log(y_i/\hat{\lambda}_i)$$

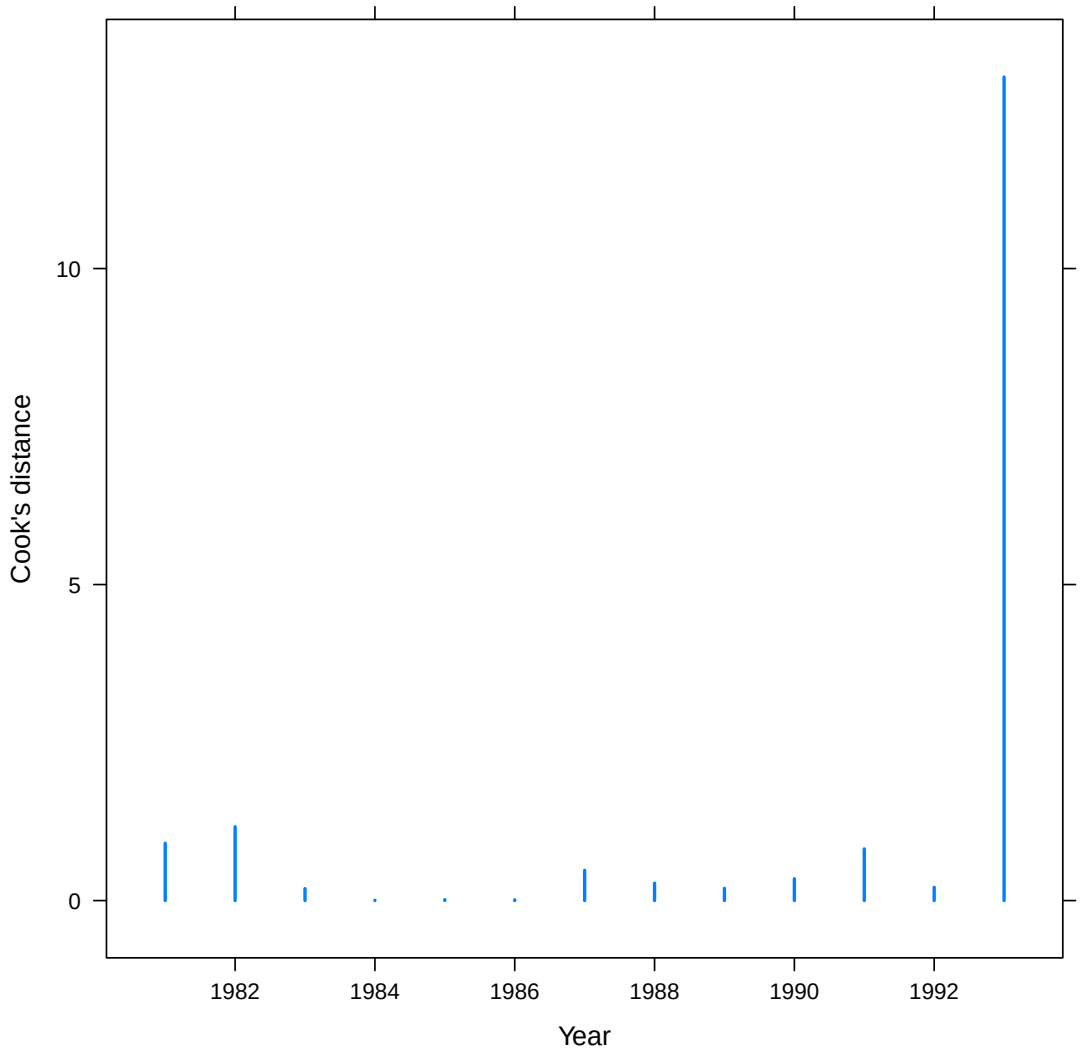
Additional residuals/diagnostics

- The concepts of leverage, leave-one-out diagnostics, Cook's distance, and Δ_β are the same as they were for logistic regression
- Recall once again that both types of residuals can be standardized by dividing by $\sqrt{1 - H_{ii}}$
- Let's take a look at what these diagnostics say about our Poisson regression fit to the Belgian AIDS data

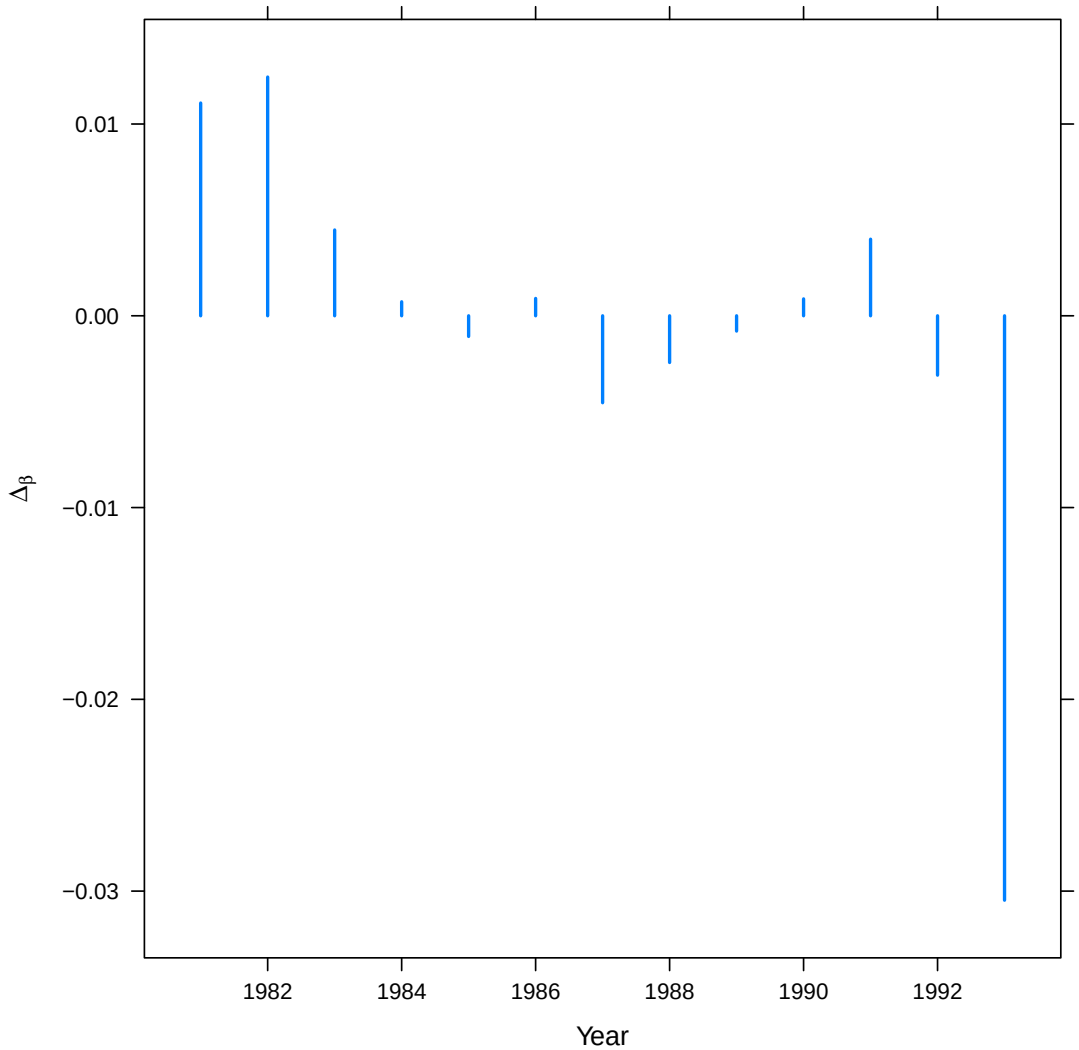
Belgian AIDS data: Leverage



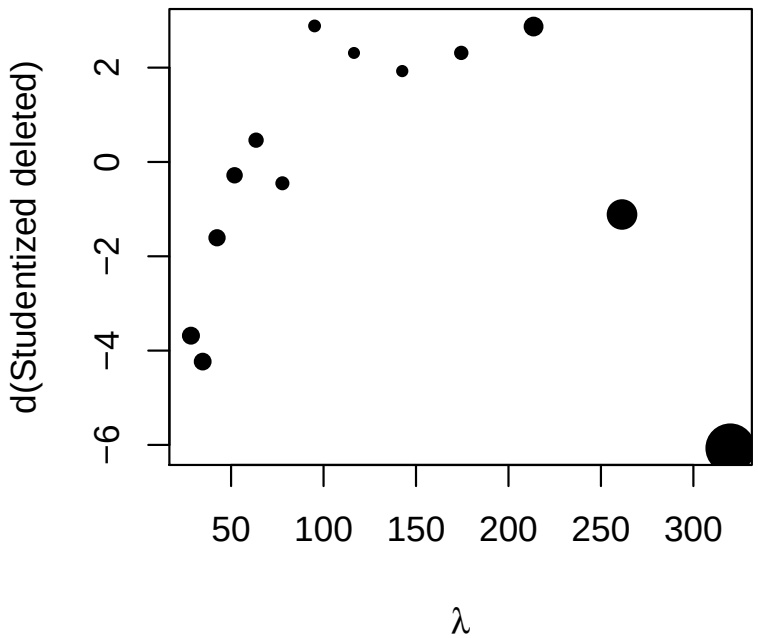
Belgian AIDS data: Influence

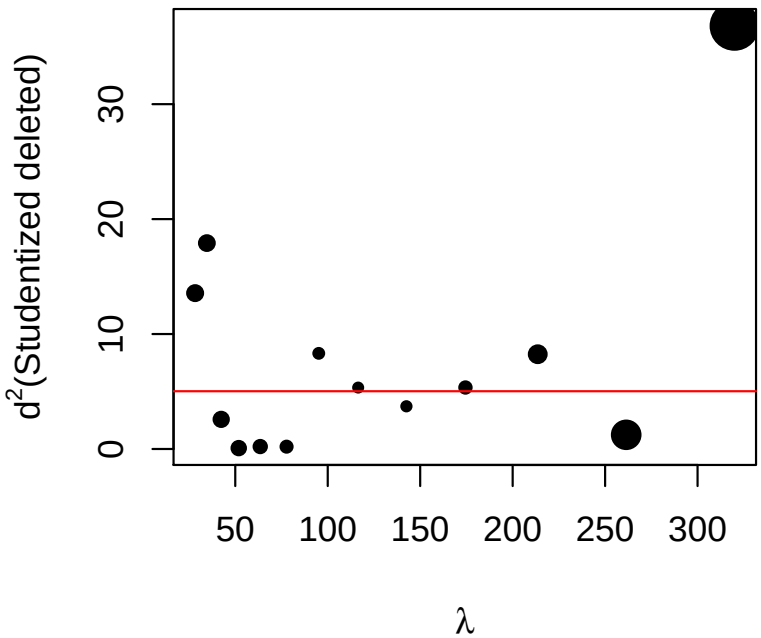


Belgian AIDS data: Δ_β (Year)



Belgian AIDS data: Residuals





Measures of predictive power

- How effective is our model at predicting the outcome?
- As with logistic regression, two measures are commonly used: reduction in squared error and deviance explained
- The reduction in squared error is

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\lambda}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- The explained deviance is

$$1 - \frac{D}{D_0}$$

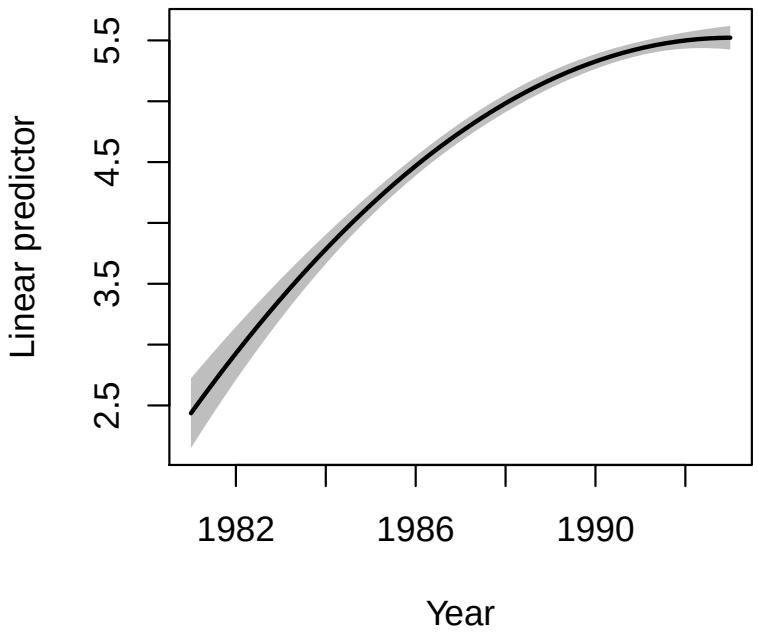
Measures of predictive power

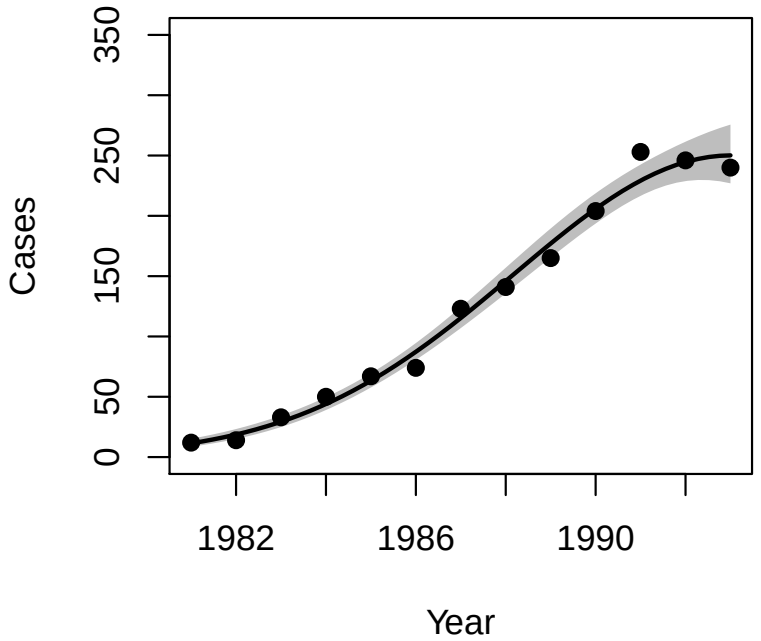
- Once again, both measures can be adjusted for number of parameters by dividing the numerator by $n - p$ and the denominator by $n - 1$
- In our example:

		R^2	R^2_{adj}	DE	DE_{adj}
1981–1993	Linear	0.880	0.869	0.907	0.899
1981–1991	Linear	0.973	0.970	0.964	0.960
1981–1993	Quadratic	0.988	0.986	0.989	0.987

- AIC also strongly favors a quadratic model (166 vs. 97)

Belgian AIDS data: Quadratic model





Poisson rates

- In more complicated models, the meaning of λ often requires additional thought
- For example, we often think of Poisson events occurring with a certain *rate*
- If this is the case, we need to be careful about specifying what we are estimating: a rate per what?
- For example, if we are modeling motor vehicle crashes, we may be estimating a rate per 1,000 population, a rate per 1,000 licensed drivers, a rate per 1,000 registered motor vehicles, or a rate per 100,000 miles traveled

British doctor study

- A kind of rate that is particularly common in epidemiological studies is a rate per person-years of follow-up
- For example, consider the classic study by Doll *et al.* in which all British male doctors were sent a questionnaire about their age and whether they smoked tobacco
- The doctors were then followed up for a number of years to see whether or not they had died from coronary heart disease

Offsets

- Suppose, then, that we wish to model $\lambda(\mathbf{x})$, the rate per 1,000 person-years of follow-up, given the explanatory variables `Age` and `Smoking`
- Now,

$$\mathrm{E}(Y_i) = t_i \lambda_i,$$

where t_i denotes the person-years of follow-up for observation i

- This implies that

$$\begin{aligned} \log(\mu_i) &= \log(t_i) + \log(\lambda_i) \\ &= \log(t_i) + \eta_i; \end{aligned}$$

thus, the usual relationship between μ_i and the linear predictor is *offset* by the amount $\log(t_i)$

Including offsets in R/SAS

- Both `R` and SAS allow you to specify an offset
- In SAS, one simply adds the option `OFFSET=` to the model statement
- Similarly, in `R`, one specifies the `offset=` option in the `glm` function
- Note: In SAS, one must compute the offset in a separate `DATA` step, while in `R`, one can submit code such as `offset=log(PersonYears/1000)`

Estimating linear combinations

- We can then estimate the rate per 1,000 person-years of follow-up for any category we choose using either the `ESTIMATE` statement in SAS or the `predict` function in `R`
- For example, with SAS's default coding of class variables, the following statement estimates the rate of CHD deaths for smokers aged 45–54:

```
ESTIMATE '45-54 smokers' Intercept 1
                         Age 0 1 0 0 0
                         Smoking 0 1;
```

- In `R`, we can set up a data frame consisting of all the linear combinations we are interested in, and then submit

```
predict(fit,df,type="response")
```

- Note: In SAS, the offset is set to zero; in `R`, you specify the offset variable

Estimated rates

- The estimated rates from our Poisson regression model:

	Smokers	Non-smokers
35–44	0.52	0.36
45–54	2.29	1.60
55–64	7.17	5.03
65–74	14.78	10.37
75–84	20.97	14.71

- Note that, by fitting a model with no interaction between age and smoking, we enforce that the rate ratio (RR) between smokers and non-smokers are the same in each age group ($0.52/0.36 = \cdots = 20.97/14.71 = 1.43$)

Rate ratios

- Rate ratios are a common way of describing the coefficients of a Poisson regression model, on a scale that is more interpretable
- This is exactly analogous to the use of odds ratios to describe logistic regression models; assume we have two observations with explanatory variable vectors $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$\begin{aligned}\frac{\hat{\lambda}_2}{\hat{\lambda}_1} &= \frac{\exp(\hat{\eta}_2)}{\exp(\hat{\eta}_1)} \\ &= \exp((\mathbf{x}_2 - \mathbf{x}_1)^T \widehat{\boldsymbol{\beta}})\end{aligned}$$

- In other words, if compare two types of individuals who are otherwise the same, but differ by one unit in x_j, the ratio of their event rates is $\exp(\hat{\beta}_j)$

Rate ratios (examples)

- So, for example, the 1.43 rate ratio we observed earlier arises from

$$RR = \exp(\hat{\beta}_{\texttt{Smoking}}) = e^{0.3545} = 1.43$$

- In the Belgian AIDS data, every five years the rate of new AIDS cases was increasing by 275%:

$$RR = \exp(5\hat{\beta}_{\texttt{Year}}) = e^{5(0.2021)} = 2.75$$

- Males 65–74 are at 6.5 times higher risk of death from CHD than males 45–54:

$$RR = \exp(\hat{\beta}_{\texttt{65--74}} - \hat{\beta}_{\texttt{45--54}}) = e^{3.3505-1.4840} = 6.5$$

Comments/connections

- Suppose we had county-level data, and were modeling occurrences of disease; should we treat the outcome as Poisson with a rate per population, or binomial with n_i the number of people in county i?
- The binomial distribution is better for small sample sizes, but if n is large and the disease is rare, it doesn't really matter; the binomial is well-approximated by the Poisson in this case
- Poisson regression is an adequate, but not ideal tool for analyzing cohort studies; if one has detailed individual-level data, one can apply the more sophisticated approaches that have been developed in the field of *survival analysis*