Poisson Regression

Patrick Breheny

April 9

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

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

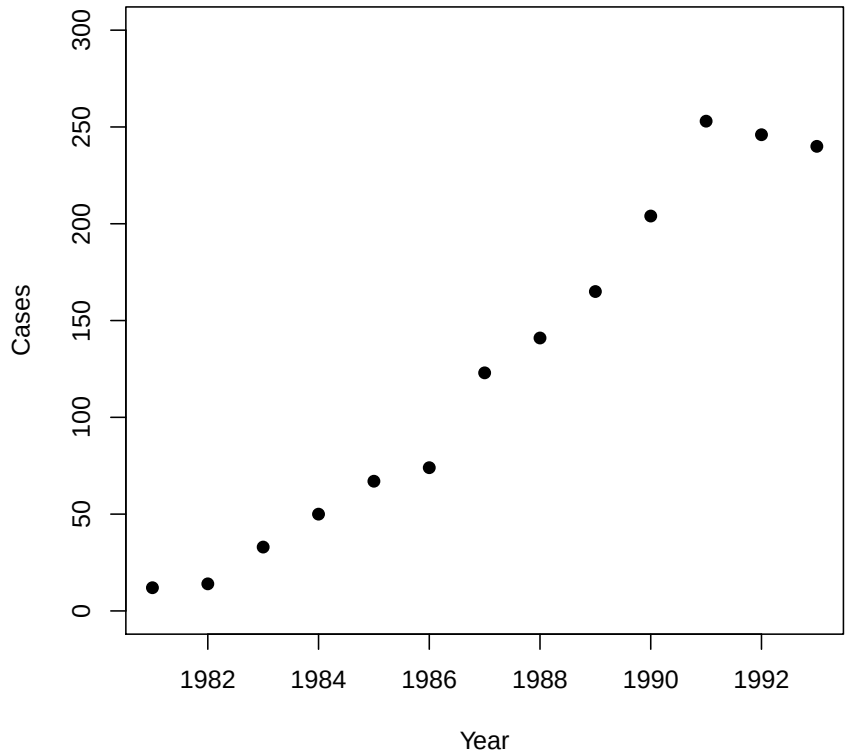
- Recall that the canonical link for the Poisson distribution is the log link
- Thus,

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

- 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

- Exponential growth models are reasonable in the early stages of an epidemic
- As we remarked back when we first started talking about GLMs, the simple linear model

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

when combined with a log link, is equivalent to fitting the exponential growth model

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

where $\beta_0 = \log(\gamma)$ and $\beta_1 = \delta$

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, data=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 again be carried out by fitting the full model (`fit`) and the reduced model (`fit0`), then submitting

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

Standard output

The standard R/SAS output is following:

	β	SE
Intercept	-397.06	15.46
Year	0.20	0.01

What does the intercept mean here?

Re-centering year at 1981

- Re-centering year so that it begins at the start of the study (1981), we obtain a meaningful intercept:

	Estimate	Std. Error
Intercept	3.34	0.07
Year	0.20	0.01

- Recall that we are modeling with a log link; the model thus estimates $e^{3.34} = 28.2$ cases in 1981
- How to interpret the coefficient for year?

Rate ratios

- Consider two hypothetical observations with different explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$; the Poisson GLM with log link implies that

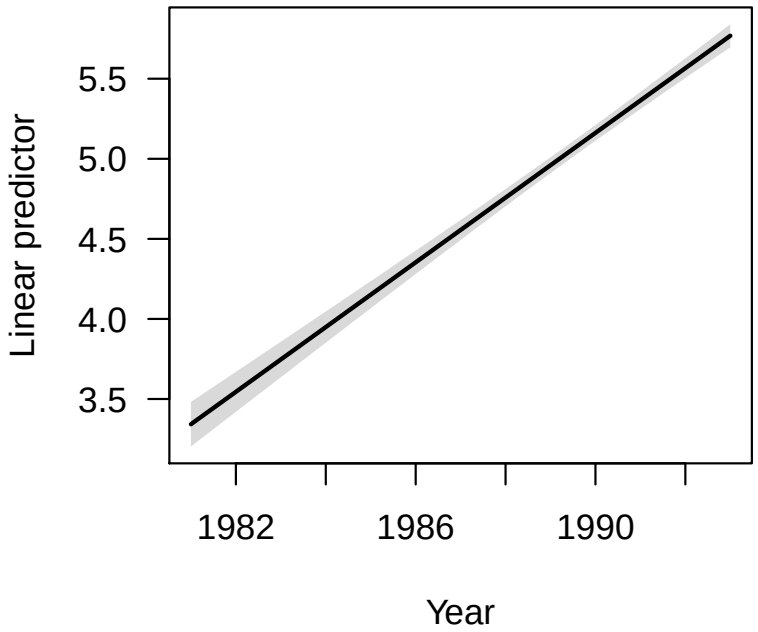
$$\begin{aligned}\frac{\lambda_2}{\lambda_1} &= \frac{\exp(\eta_2)}{\exp(\eta_1)} \\ &= \exp((\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta})\end{aligned}$$

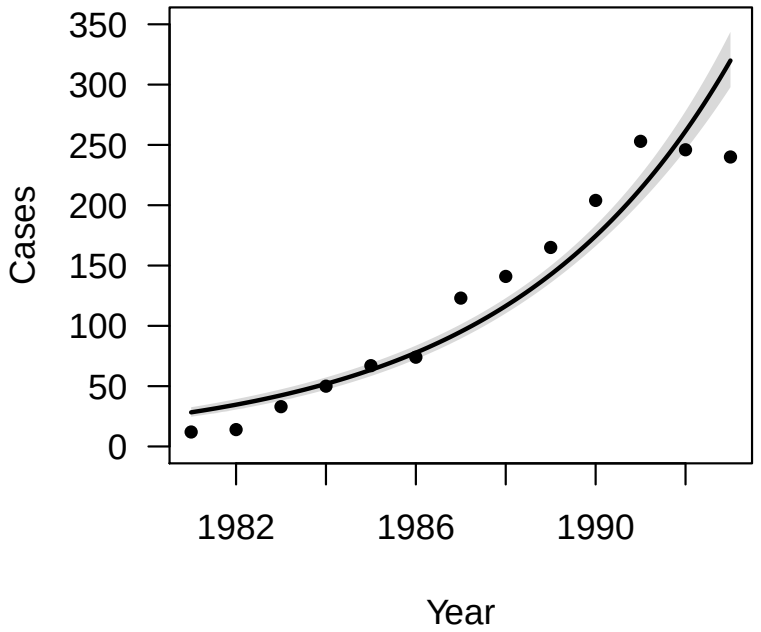
- In particular, if variable j changes by an amount δ_j, the *rate ratio* λ_2/λ_1 is $\exp(\delta_j \beta_j)$
- Rate ratios (RR) are a common way of describing the coefficients of a Poisson regression model, putting them on a scale that is more interpretable, analogous to the use of odds ratios in logistic regression models

Rate ratios: Examples

- So, our regression coefficient of 0.20 implies that the rate ratio is $e^{0.20} = 1.2$; the number of AIDS cases in Belgium increased by 20% each year over the time span 1981-1993
- Another way of putting it is that $e^{5(0.20)} = 2.7$; the number of AIDS cases increased by 170% every five years
- Or yet another way of putting it, $e^{3.5(0.20)} = 2$; the number of AIDS cases doubled every 3.5 years

Visualizing the model





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 our definition of deviance
- Previously, we have taken deviance to mean -2ℓ; a broader definition is

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

  where $\ell_{\max}$ is the maximum possible log-likelihood for the observed data, given the distribution specified by the model
- Here, deviance may be interpreted as the gap between a model's fit to the data and the fit of an ideal model for which $\hat{\mu}_i = y_i$ for all observations
- This detail was not relevant to our earlier uses of deviance, as for the Bernoulli and normal distributions, $\ell_{\max} = 0$

Deviance residuals

- This is not the case for the Poisson distribution, however
- For the Poisson distribution,

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

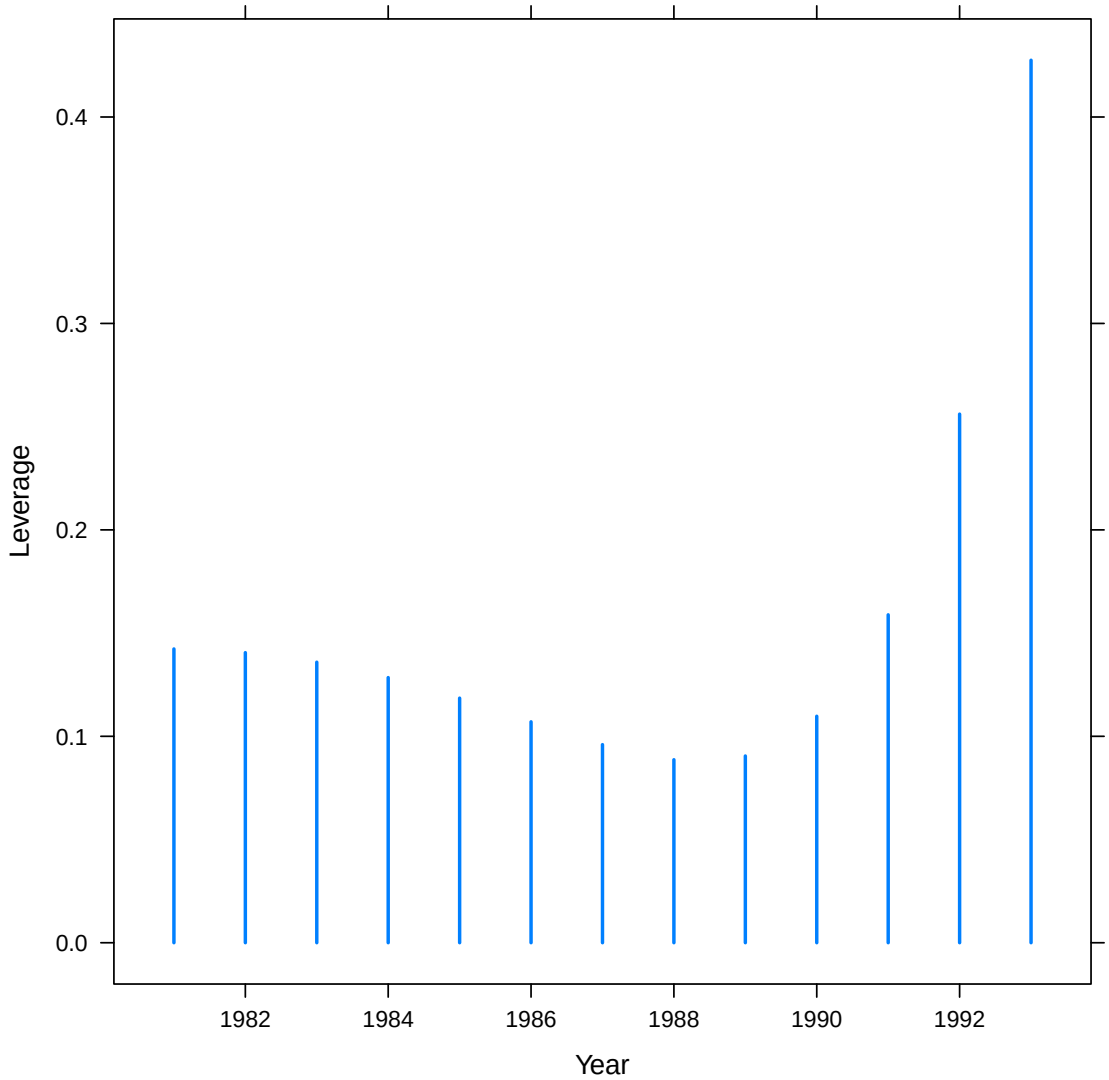
  where you may recall that s_i was the sign of $y_i - \hat{\lambda}_i$
- Note the advantage of our new deviance definition: it allows all the $y_i!$ terms to cancel out
- The deviance is $D = \sum_i d_i^2$, although if the model has an intercept, then $\sum_i y_i = \sum_i \hat{\lambda}_i$, and the deviance simplifies to

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

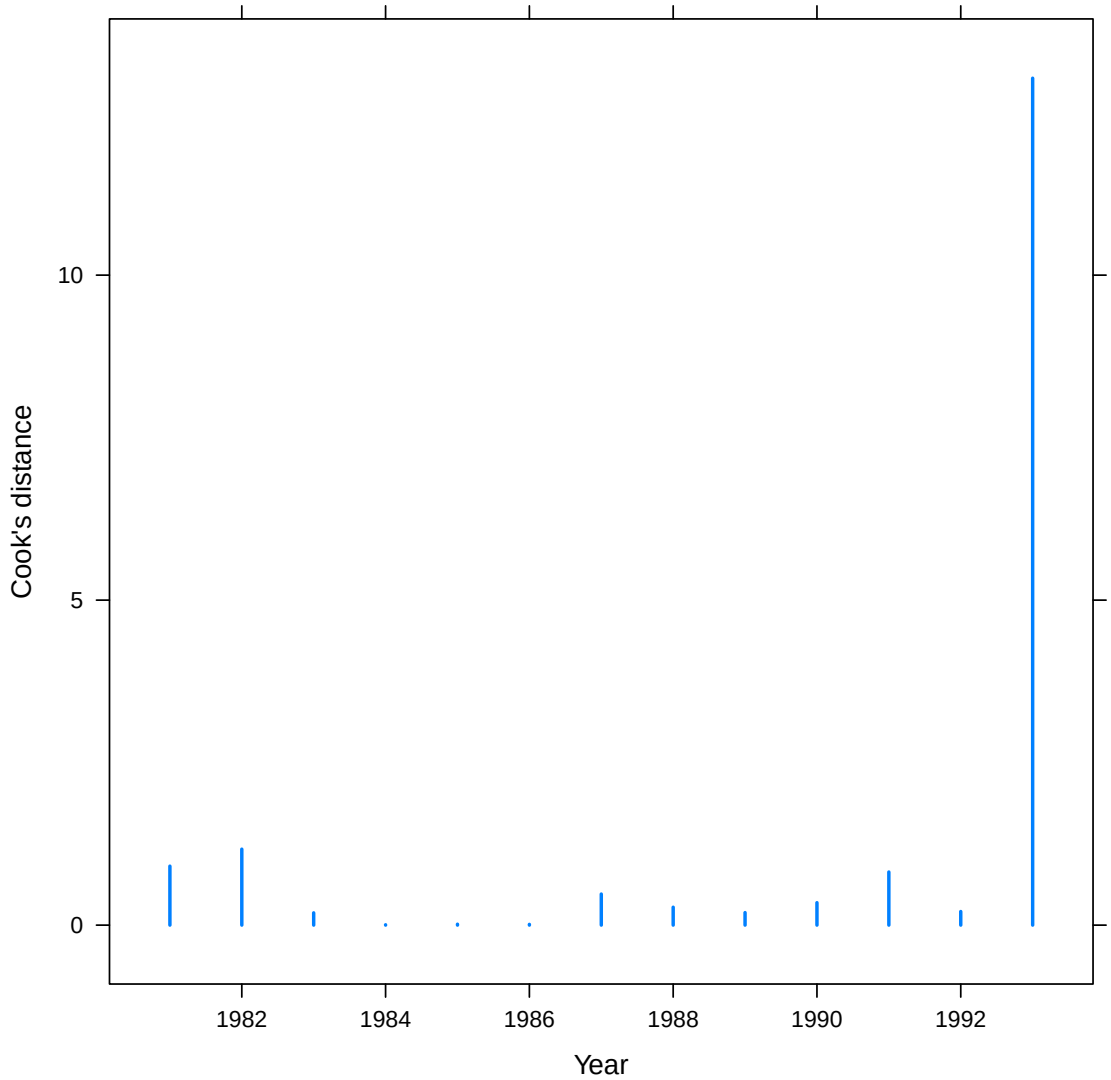
Additional residuals/diagnostics

- 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

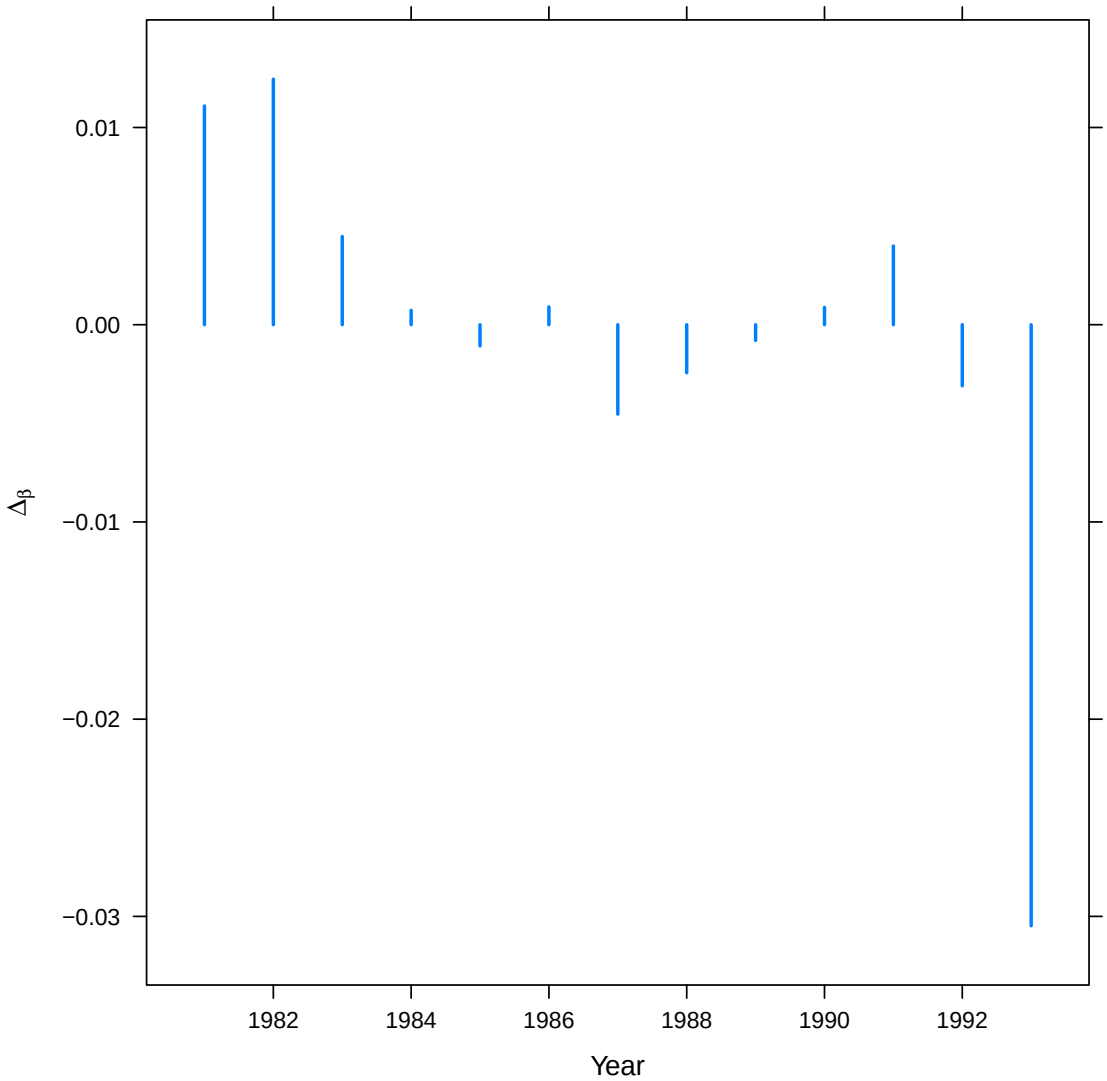
Belgian AIDS data: Leverage



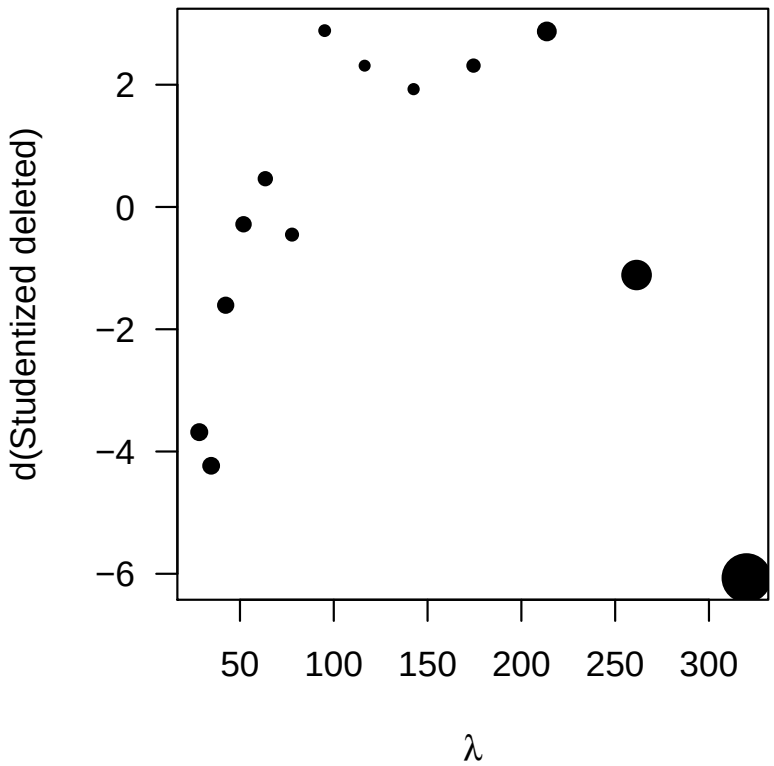
Belgian AIDS data: Influence

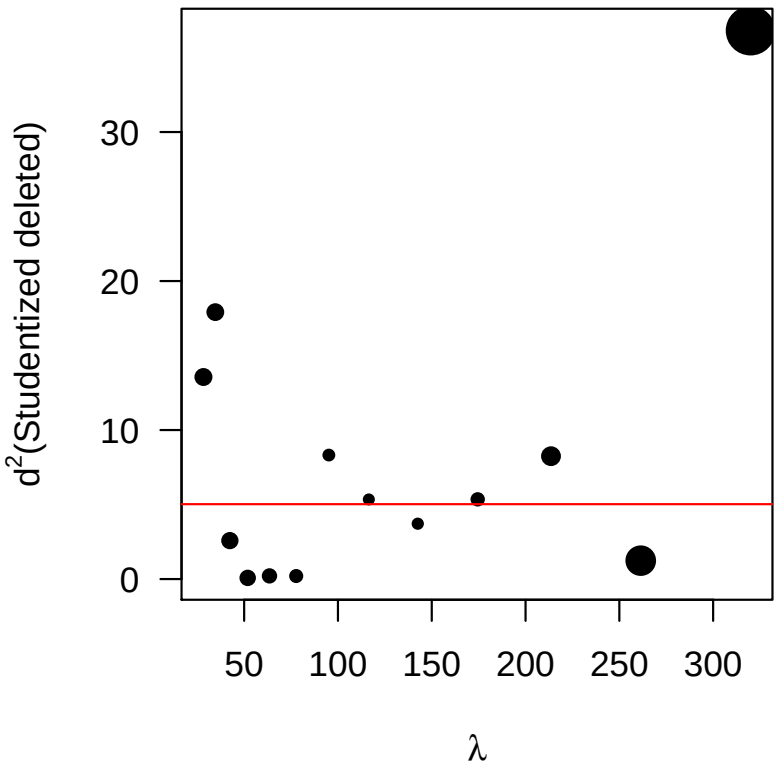


Belgian AIDS data: Δ_β (Year)



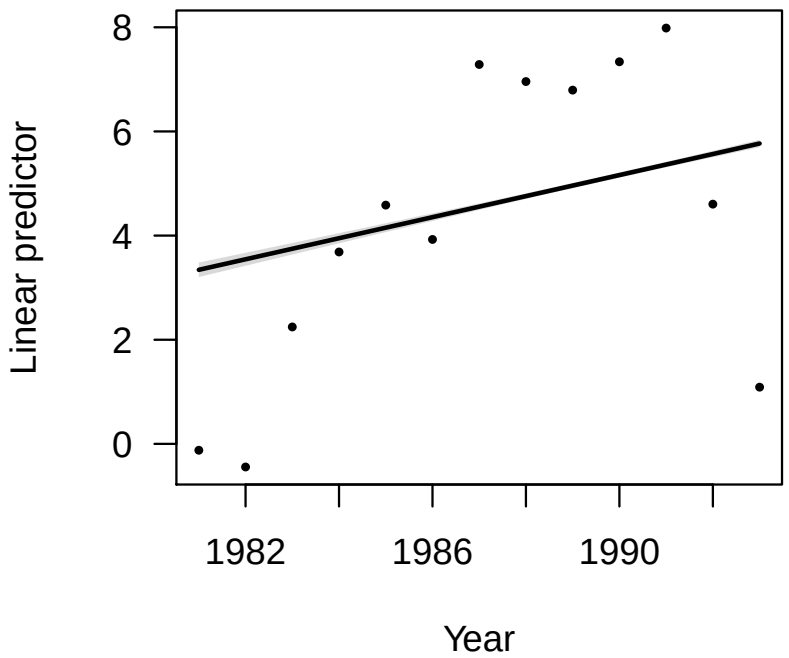
Belgian AIDS data: Residuals





Considering a quadratic model

- All of these plots indicate problems – our model fits the data from 1992 and 1993 poorly, and this has a fairly large impact
- A plot of the partial residuals suggests fitting a quadratic model:



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

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

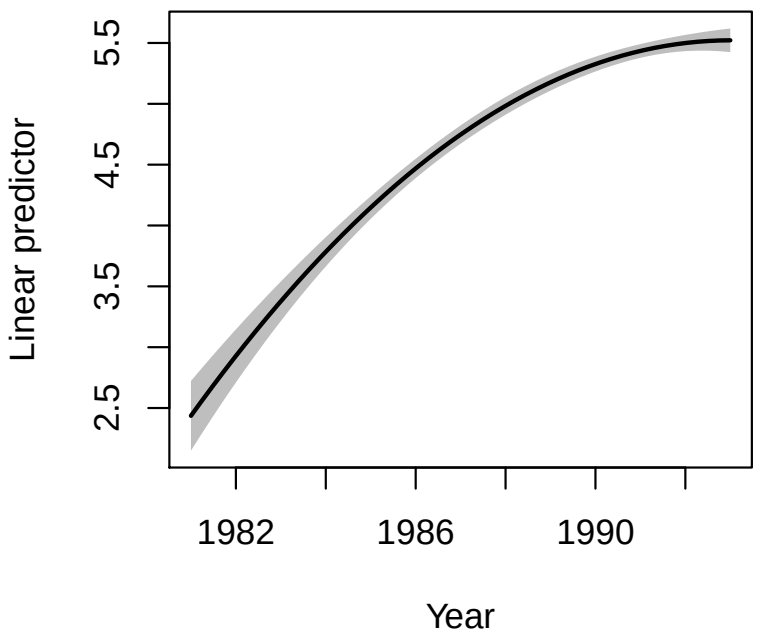
Measures of predictive power

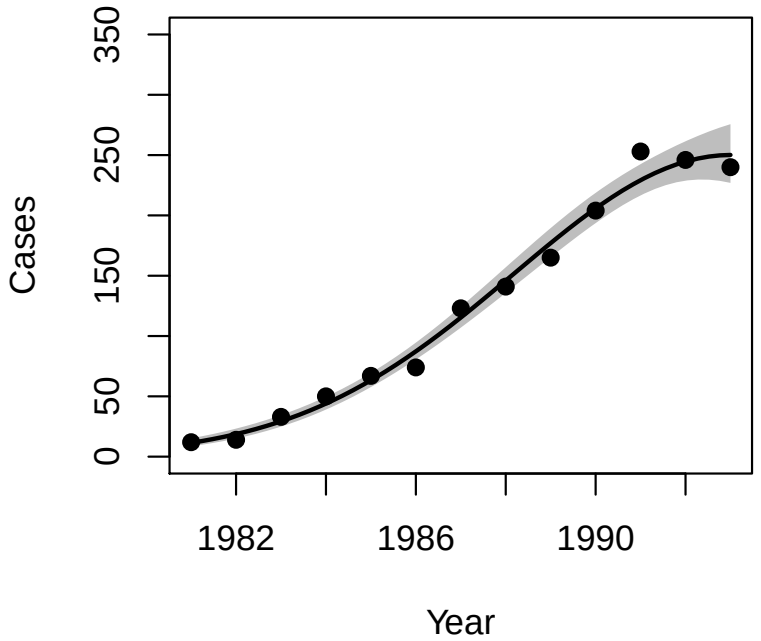
- Once again, both measures can be adjusted for number of parameters by dividing the numerator by $n - p$ and the denominator by $n - 1$
- In our example:

		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

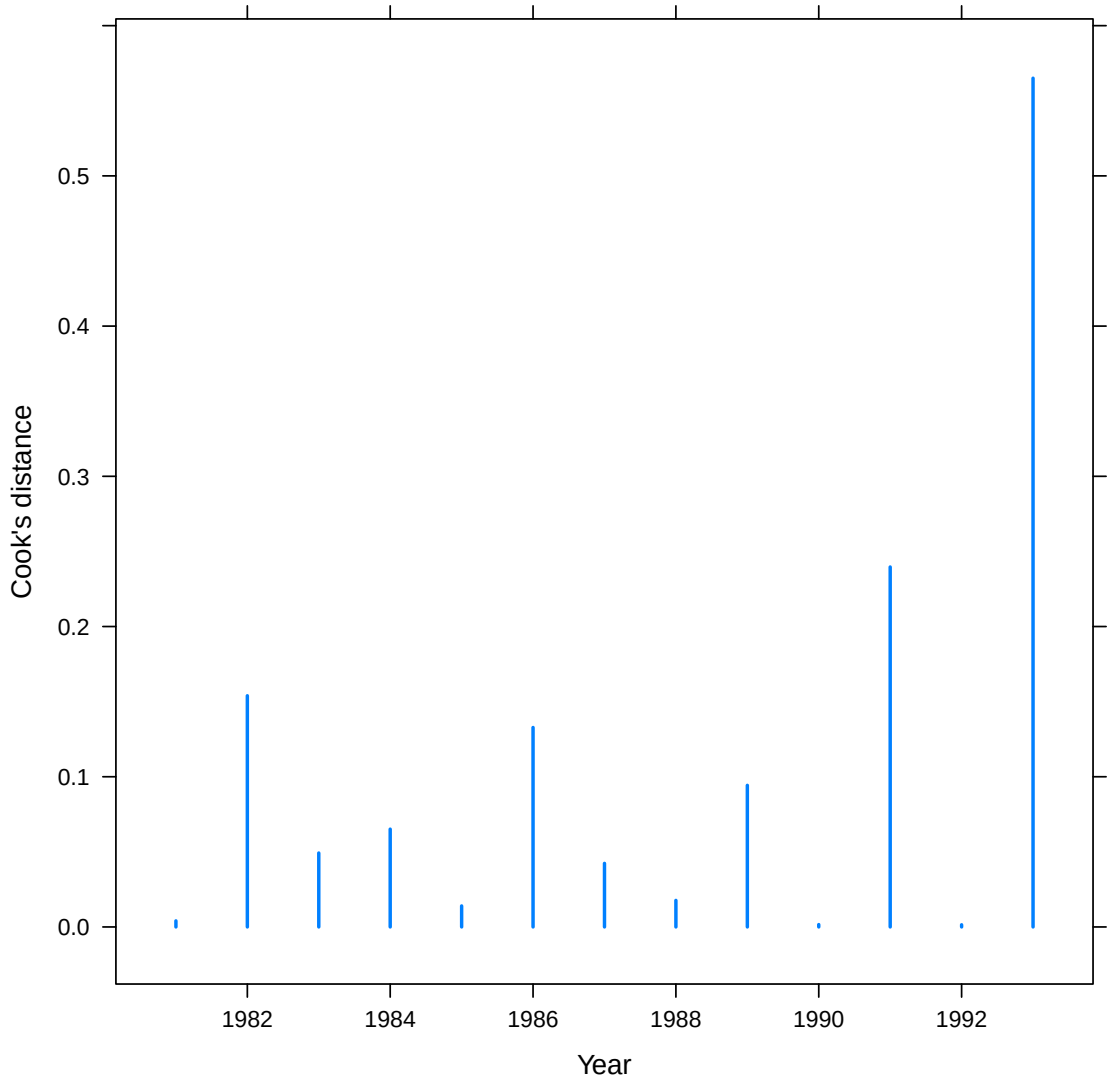
- AIC also strongly favors a quadratic model (166 vs. 97)

Belgian AIDS data: Quadratic model

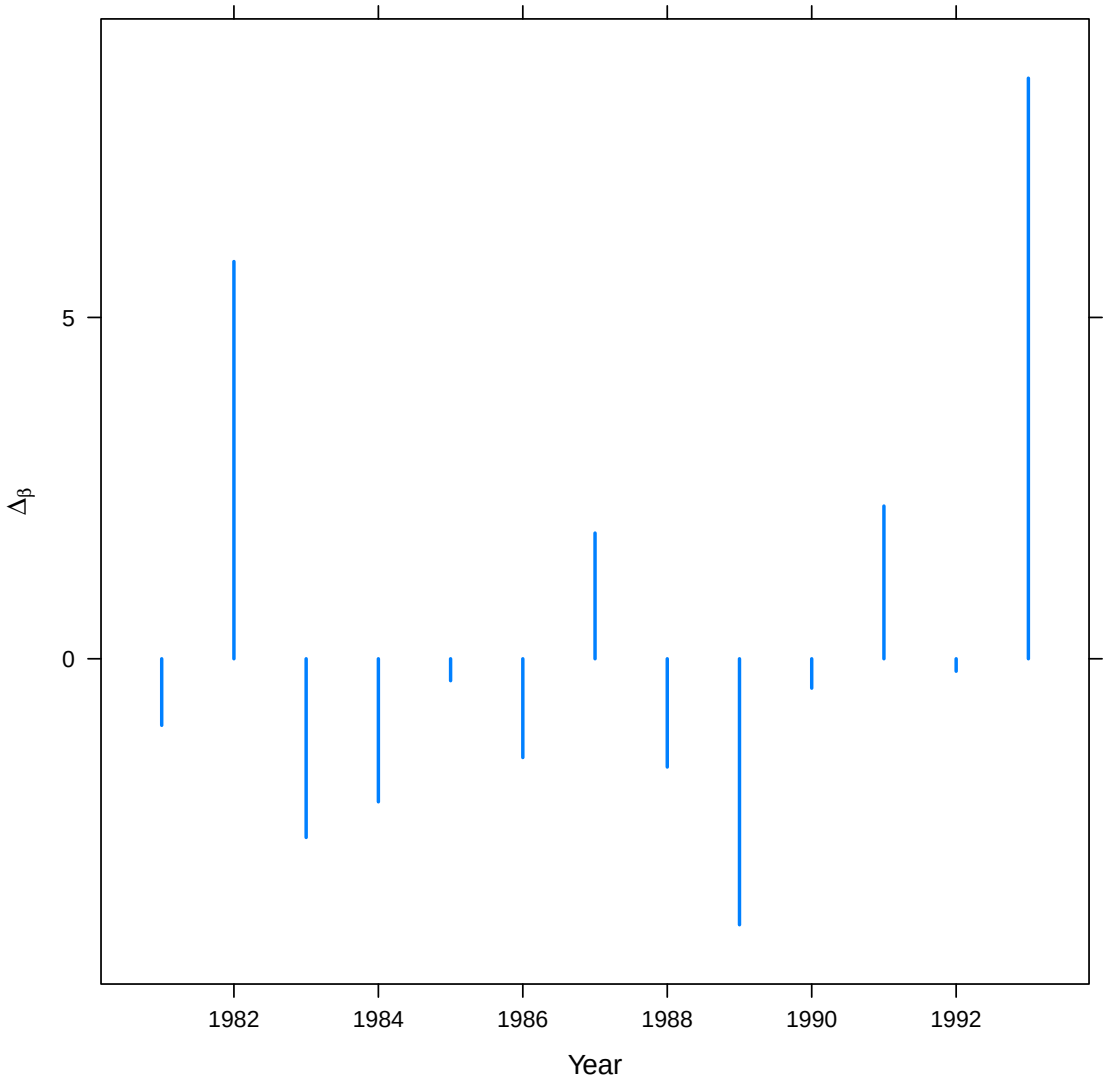




Belgian AIDS data: Influence for quadratic model



Belgian AIDS data: Δ_β (Year) for quadratic model



Belgian AIDS data: Residuals for quadratic model