Generalized Linear Models

Simon Wood
Mathematical Sciences, University of Bath, U.K.

Generalized linear model

- The linear model $y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ can be written

$$\mu_i = \mathbf{X}_i\boldsymbol{\beta} \qquad y_i \sim N(\mu_i, \sigma^2).$$

  where $\mu_i = E(y_i)$.
- A *Generalized linear model* (GLM) extends this somewhat

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} \qquad y_i \sim \text{EF}(\mu_i, \phi)$$

  - g is any smooth monotonic *link function*.
  - $\text{EF}(\mu_i, \phi)$ is any *exponential family distribution* (e.g. Normal, gamma, Poisson, binomial, Tweedie, etc.)
  - ϕ is a known or unknown *scale parameter*
  - $\mathbf{X}\boldsymbol{\beta}$ $(= \boldsymbol{\eta})$ is the *linear predictor*

The link function, g

- Common link functions are log, square root and logit ($\log\{\mu_i/(1-\mu_i)\}$).
- g acts a *little bit* like the data transformations used before GLMs. However note:
  - The link function transforms $\mathbb{E}(y_i)$.
  - The link function **does not** transform y_i itself.
- So, with a GLM we can transform the systematic part of a model, without changing the distribution of the random variability.

The exponential family

- A distribution is in the exponential family if its probability (density) function can be written in a particular general form.
- For our purposes, what matters is that if y is from an exponential family distributions, then we can write:

$$\text{var}(y) = V(\mu)\phi$$

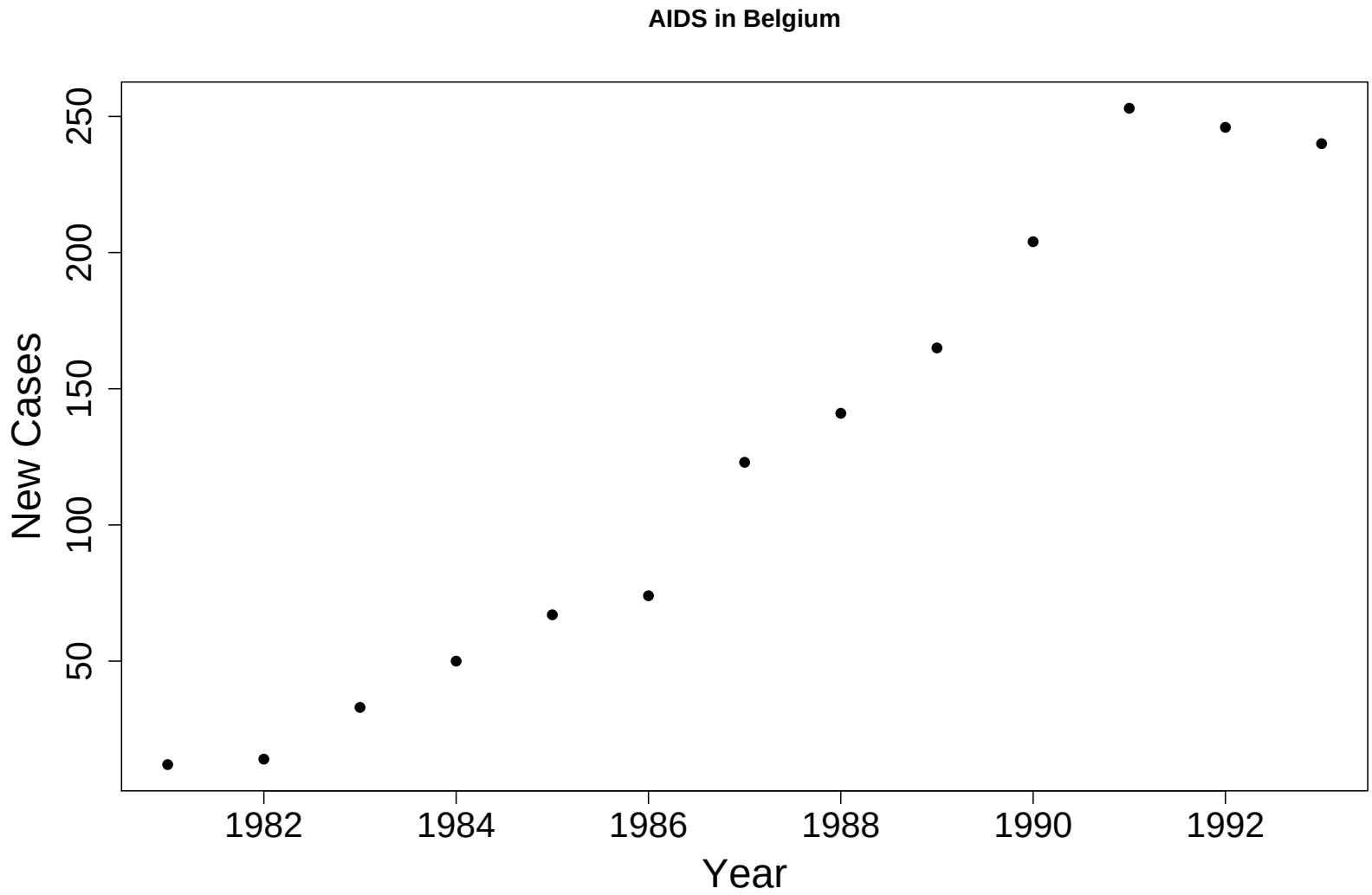
  where V is a known *variance function* of $\mu = \mathbb{E}(y)$, and ϕ is a **scale parameter** (known or unknown).
- Actually GLM theory can be made to work based only on knowledge of V, without needing to know the full distribution of y, using *quasi-likelihood* theory.

GLMs: scope

Generalized linear models include many familiar model types, for example:

- Linear models. Identity link, normal distribution.
- Models for analysis of contingency tables. Log link, Poisson distribution.
- Logistic regression. ‘logit’ or ‘probit’ link, binomial distribution.

Example: AIDS in Belgium



Example: AIDS rate model

- A simple model for these data might be

$$\mathbb{E}(y_i) = N_0 e^{\beta_1 t_i} \qquad y_i \sim \text{Poi}$$

  - y_i is new cases in year t_i; N_0 is number of new cases in 1980.
  - Model is for exponential increase of the expected rate.
  - Observed number of cases, follows a Poisson distribution.

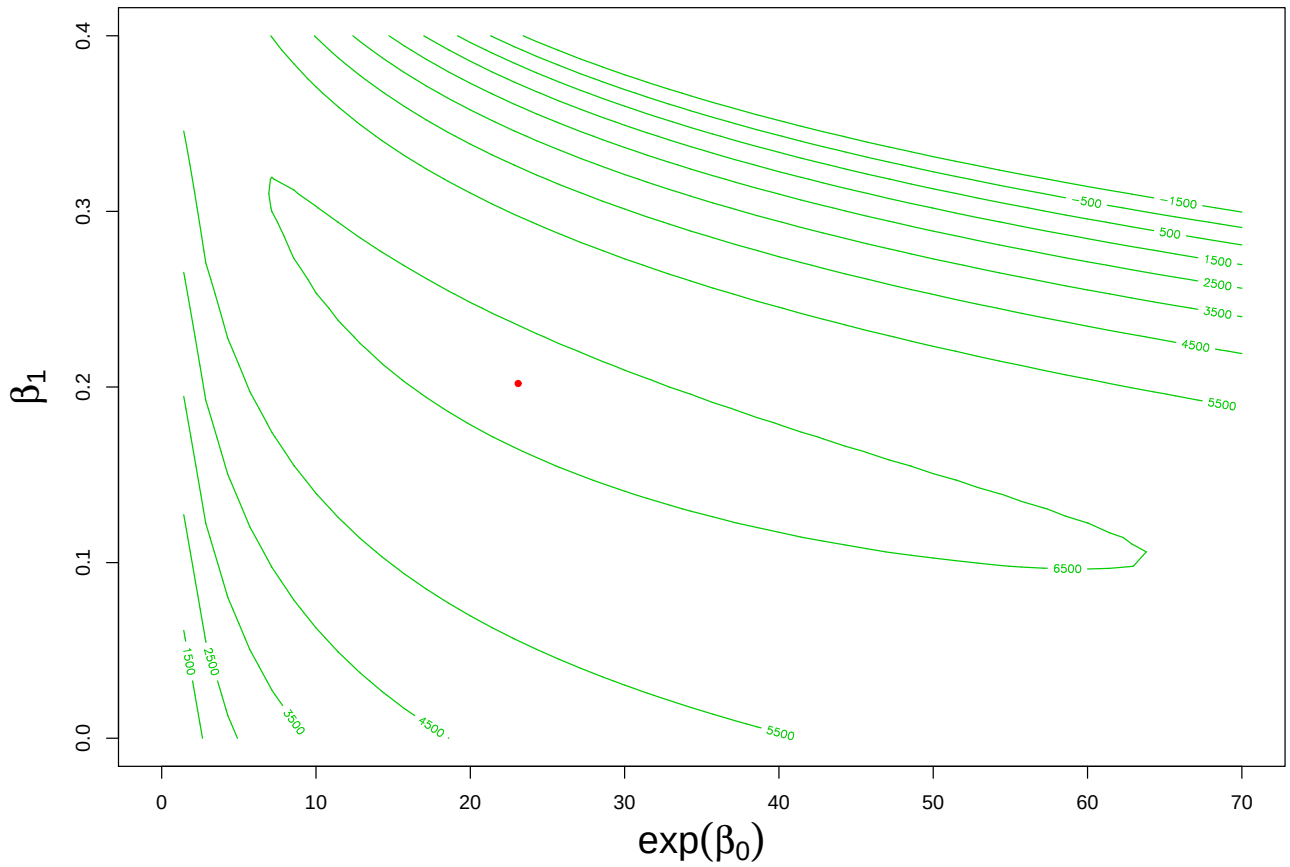
- The model is non-linear, but taking logs yields

$$\begin{aligned} \log\left(\mathbb{E}(y_i)\right) &= \log(N_0) + \beta_1 t_i \\ &= \beta_0 + \beta_1 t_i, \qquad y_i \sim \text{Poi} \end{aligned}$$

  i.e. a GLM with a log link ($\beta_0 \equiv \log(N_0)$).

GLM estimation

Model estimation is by maximum likelihood, via a Newton type method. e.g. for the AIDS model the log-likelihood function looks like this . . .



IRLS

- For GLMs, MLE by a Newton type method can be expressed as an Iteratively Re-weighted Least Squares scheme. . .
- Initialize $\hat{\eta}_i = g(y_i)$, then iterate the following steps.
  1. Form pseudodata $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha_i + \hat{\eta}_i$ and iterative weights, $w_i = \alpha_i / \left\{ V(\hat{\mu}_i) g'(\hat{\mu}_i)^2 \right\}$.
  2. Minimize the weighted sum of squares $\sum_i w_i (z_i - \mathbf{X}_i \boldsymbol{\beta})^2$ w.r.t. $\boldsymbol{\beta}$ to obtain a new $\hat{\boldsymbol{\beta}}$, and hence new $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$.
- $\alpha_i = 1 + (y_i - \hat{\mu}_i)(V_i'/V_i + g_i''/g_i')$ gives Newton's method.
- $\alpha_i = 1$ gives *Fisher scoring*, where the expected Hessian of the likelihood replaces the actual Hessian in Newton's method.
- Newton convergences faster. Every EF has a *canonical link* for which the Newton $\equiv$ Fisher.
- At convergence $\hat{\boldsymbol{\beta}}$ is the MLE (both methods!).

Distribution of $\hat{\beta}$

- In the large sample limit, by MLE theory (or from the weighted least squares),

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X})^{-1}\phi).$$

- Hence, CIs for any β_i can be calculated.
- Often ϕ is known. e.g. $\phi = 1$ for Poisson or binomial.
- If ϕ is unknown, can use a *Pearson estimate*:

$$\hat{\phi} = \sum_i w_i(z_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2/(n - \dim(\boldsymbol{\beta}))$$

(then need to use $t_{n-\dim(\boldsymbol{\beta})}$ distribution for CI's).

Deviance

- It is useful to have a quantity for GLMs which behaves like the residual sum of squares of a linear model. This is the *deviance*.
- We can write the model log likelihood, $l(\boldsymbol{\beta})$, as a function of the $\boldsymbol{\mu}$: $l(\boldsymbol{\mu})$. Then the *deviance* is

$$D = 2\left\{l(\mathbf{y}) - l(\hat{\boldsymbol{\mu}})\right\}\phi$$

- It turns out that D can be evaluated without knowing ϕ, but for hypothesis testing we need the *scaled deviance* $D^* = D/\phi$. (When $\phi = 1$, $D^* = D$).

Properties of deviance

- The deviance reduces as the model fit improves.
- If the model exactly fits the data then the deviance is zero.
- As a rough approximation

$$D^* \sim \chi^2_{n-\dim(\boldsymbol{\beta})}$$

  if the model is correct. Approximation can be good in some cases and is exact for the strictly linear model.
- This suggests an alternative scale parameter estimator

$$\hat{\phi} = D/\{n - \dim(\boldsymbol{\beta})\}.$$

  (Since $E(\chi^2_p) = p$ and $D^* = D/\phi$.)

Model Comparison

- Nested GLMs 0 and 1, with p_0 and p_1 parameters, can be compared using a generalized likelihood ratio test...
- In terms of the scaled deviance, if model 0 is correct then $D_0^* - D_1^* \sim \chi^2_{p_1-p_0}$.
  1. If $\phi = 1$ this means that under model 0: $D_0 - D_1 \sim \chi^2_{p_1-p_0}$.
  2. If ϕ is unknown, then the GLRT leads to the approximate result that, under model 0

$$\frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \sim F_{p_1-p_0, n-p_1}.$$

- AIC can be used if we want the best model for prediction, rather than the simplest model supportable by the data.

Residuals for GLMs

- For GLMs we need to check the assumptions that the data are **independent** and have the assumed **mean-variance** relationship, and are consistent with the assumed **distribution**.
- From the raw residuals $\hat{\epsilon}_i = y_i - \mu_i$ it is very difficult to check the mean variance relationship or distribution.
- We therefore *standardize* the residuals, so that they have approximately constant variance, and behave something like residuals for an ordinary linear model.

Pearson residuals

- Pearson residuals are obtained by dividing the raw residuals by their scaled standard deviation, according to the model

$$\epsilon_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- Hence, if the model mean variance relationship is OK, the variability of these residuals should appear to be fairly constant, when they are plotted against fitted values or predictors.
- Pearson residuals are still skewed, if the distribution is skewed.

Deviance residuals

- For a linear model the *residual sum of squares* is the sum of the *squared residuals*.
- For a GLM the deviance can be written as the sum of deviances for each datum:

$$D = \sum d_i$$

- Since the deviance is supposed to behave a bit like the RSS, then by analogy we can view $\sqrt{d_i}$ as a residual.
- Specifically $\epsilon_i^d = \text{sign}(y_i - \mu_i)\sqrt{d_i}$, behave quite like residuals from a linear model.

`glm` in R

GLMs are fitted using `glm`, which functions much like `lm`

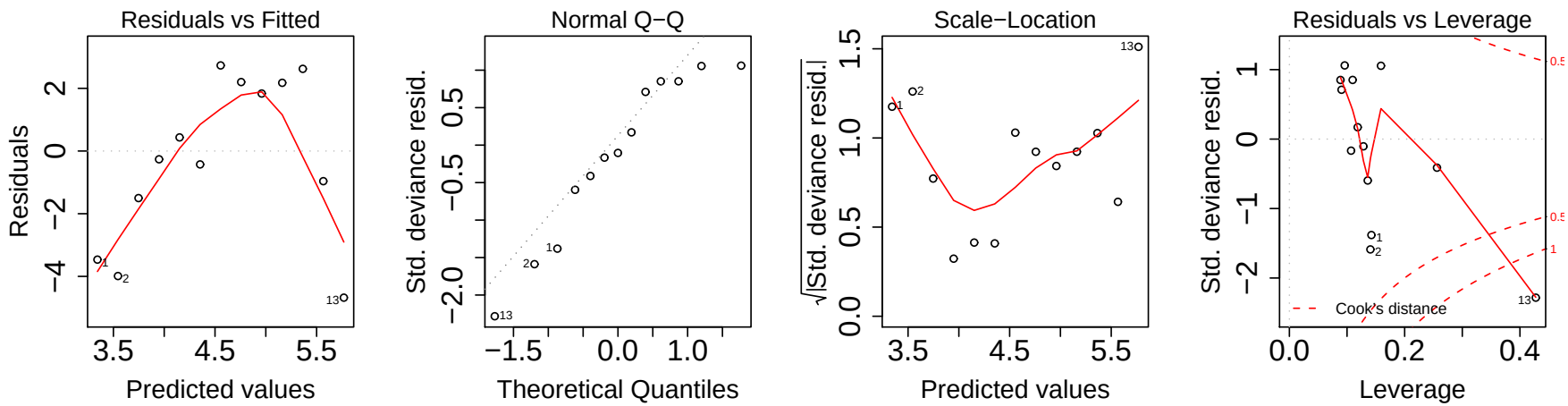
- A model formula specifies the response variable on the left, and the structure of the linear predictor on the right.
- A `data` argument is usually supplied, containing the variables referred to by the formula.
- `glm` returns a fitted model object.

But we must now specify a distribution and link.

- The `family` argument achieves this.
- e.g. `glm(...,family=poisson(link=log))` would fit a model with a log link assuming a Poisson response variable.

AIDS model example

```
belg.aids <- data.frame(cases=c(12,14,33,50,67,74,123,
                141,165,204,253,246,240),year=1:13)
am1 <- glm(cases ~ year,data=belg.aids,
           family=poisson(link=log))
plot(am1)
```

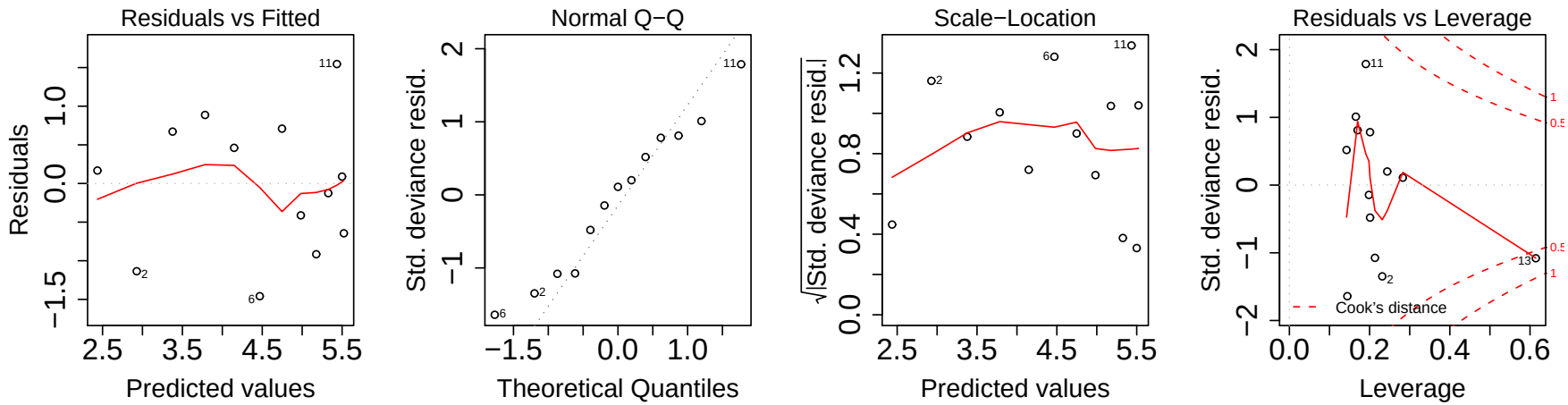


. . . clear trend in the residual mean + some overly influential points.

AIDS model example II

Try a quadratic time dependence?

```
am2 <- glm(cases ~ year+I(year^2),data=belg.aids,
           family=poisson(link=log))
plot(am2)
```



. . . much better.

AIDS example III

Now, examine the fitted model, first with the default print method

```
> am2

Call:  glm(formula=cases~year+I(year^2),
           family=poisson(link=log),data=belg.aids)

Coefficients:
(Intercept)         year    I(year^2)
    1.90146      0.55600     -0.02135

Degrees of Freedom: 12 Total (i.e. Null);  10 Residual
Null Deviance:      872.2
Residual Deviance: 9.24         AIC: 96.92
```

summary.glm (edited)

```
> summary(am2)
...
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.45903  -0.64491   0.08927   0.67117   1.54596

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.901459   0.186877  10.175  < 2e-16 ***
year         0.556003   0.045780  12.145  < 2e-16 ***
I(year^2)   -0.021346   0.002659  -8.029 9.82e-16 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 872.2058  on 12  degrees of freedom
Residual deviance:   9.2402  on 10  degrees of freedom
AIC: 96.924
```

Hypothesis testing

We can also use a GLRT/ analysis of deviance to test the null hypothesis that am1 is correct, against the alternative that am2 is ...

```
> anova(am1,am2,test="Chisq") ## NOT doing ANOVA!
Analysis of Deviance Table

Model 1: cases ~ year
Model 2: cases ~ year + I(year^2)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        11     80.686
2        10      9.240  1   71.446 2.849e-17
```

... very strong evidence against am1.

Further model improvement?

Would a cubic term be an improvement?

```
>  ## NOT doing ANOVA!
Analysis of Deviance Table

Model 1: cases ~ year + I(year^2)
Model 2: cases ~ year + I(year^2) + I(year^3)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        10     9.2402
2         9     9.0081  1   0.2321    0.6300
```

. . . no evidence that it would.

AIC comparison

```
> AIC(am1,am2,am3)
     df        AIC
am1   2 166.36982
am2   3  96.92358
am3   4  98.69148
```

So, both hypothesis testing and AIC agree that the quadratic model, am2 is the most appropriate.

predict

- `predict` method functions are the standard way of obtaining predictions from a fitted model object.
- The predictor variable values at which to predict are supplied in a `newdata` dataframe. If absent the values used for fitting are employed.
- The following predicts from `am2`, with standard errors.

```
year <- seq(1,13,length=100)
fv <- predict(am2,newdata=data.frame(year=year),se=TRUE)
```

Now we can plot the data, fitted curve, and standard error bands:

```
plot(belg.aids$year+1980,belg.aids$cases)  # data
lines(year+1980,exp(fv$fit),col=2)         # fit
lines(year+1980,exp(fv$fit+2*fv$se),col=3) # upper c.l.
lines(year+1980,exp(fv$fit-2*fv$se),col=3) # lower c.l.
```

Fitted AIDS model