# Lecture 3

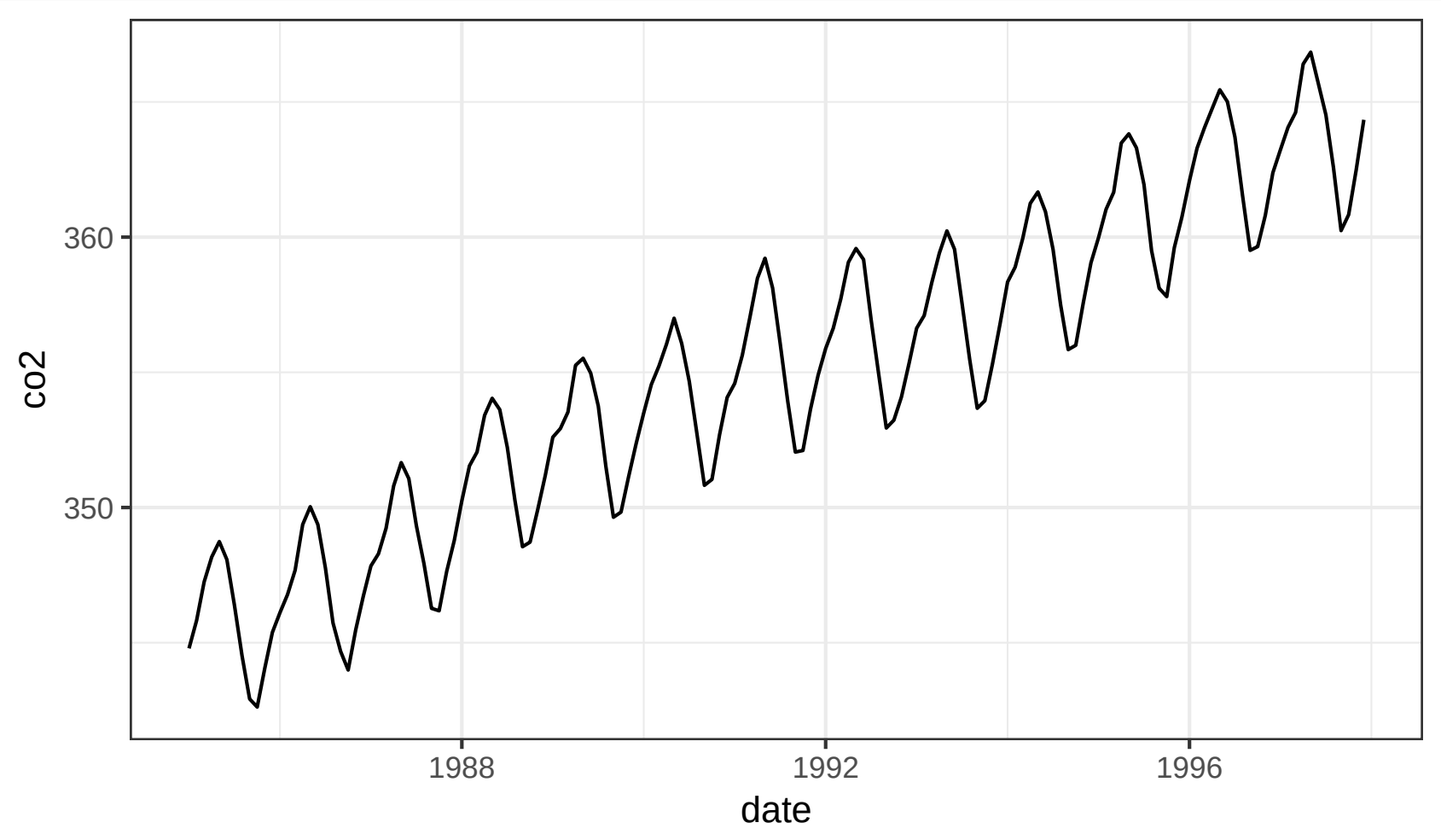
## Residual Analysis + Generalized Linear Models

Colin Rundel

1/23/2017

# Residual Analysis

# Atmospheric $CO_2$ (ppm) from Mauna Loa

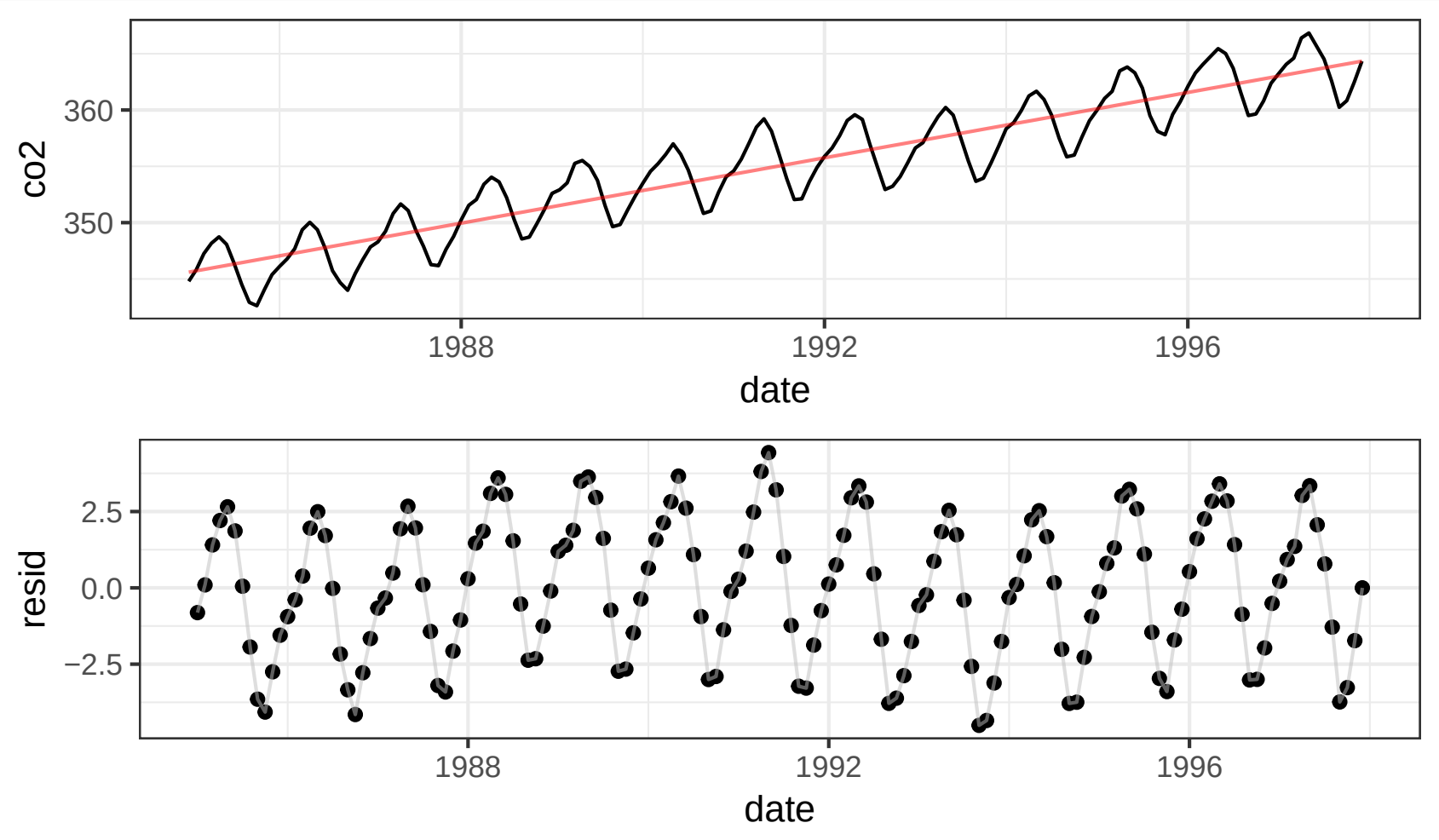


# Where to start?

Well, it looks like stuff is going up on average ...

# Where to start?

Well, it looks like stuff is going up on average ...

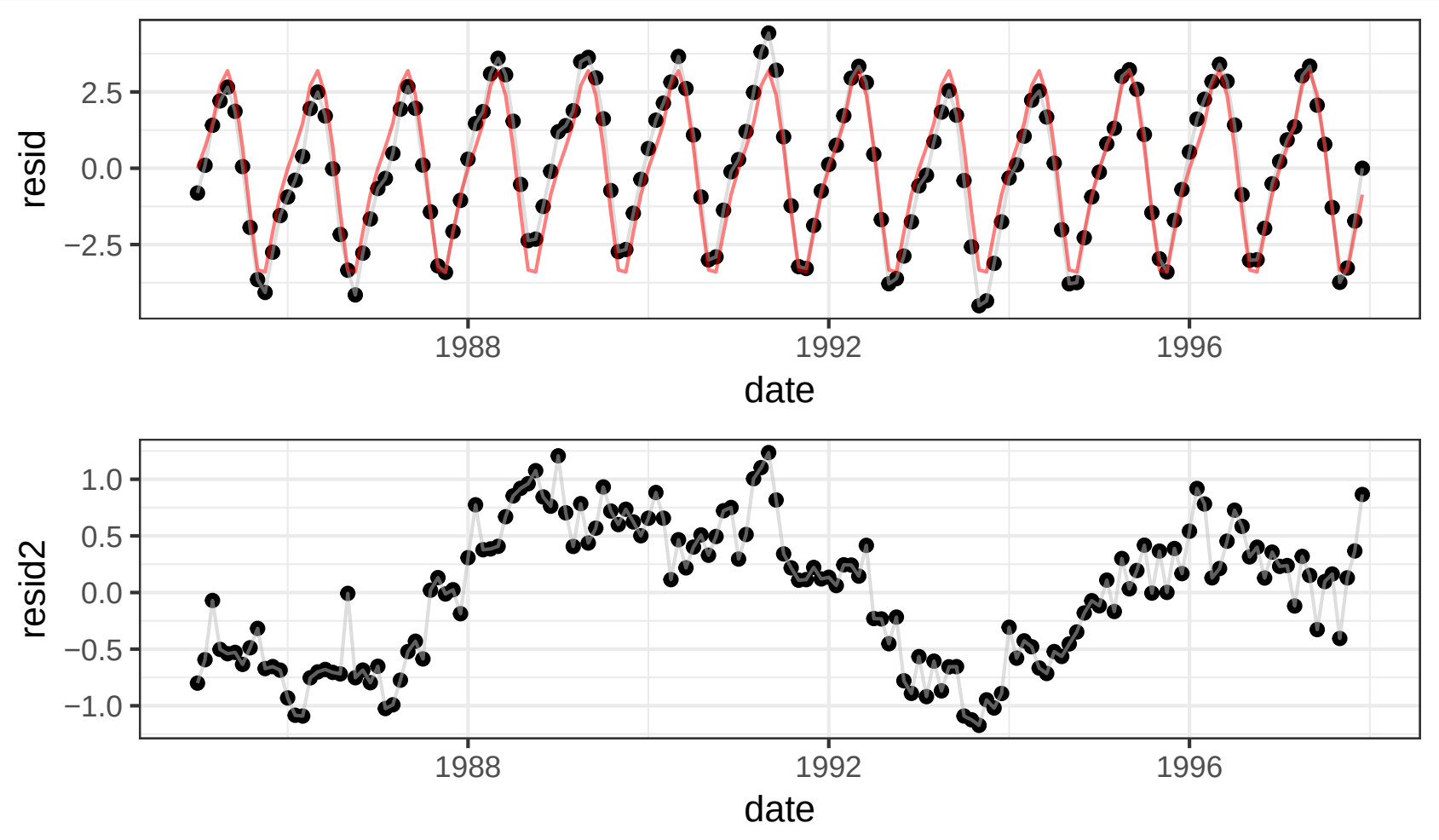


# and then?

Well there is some periodicity lets add the month ...

# and then?

Well there is some periodicity lets add the month ...

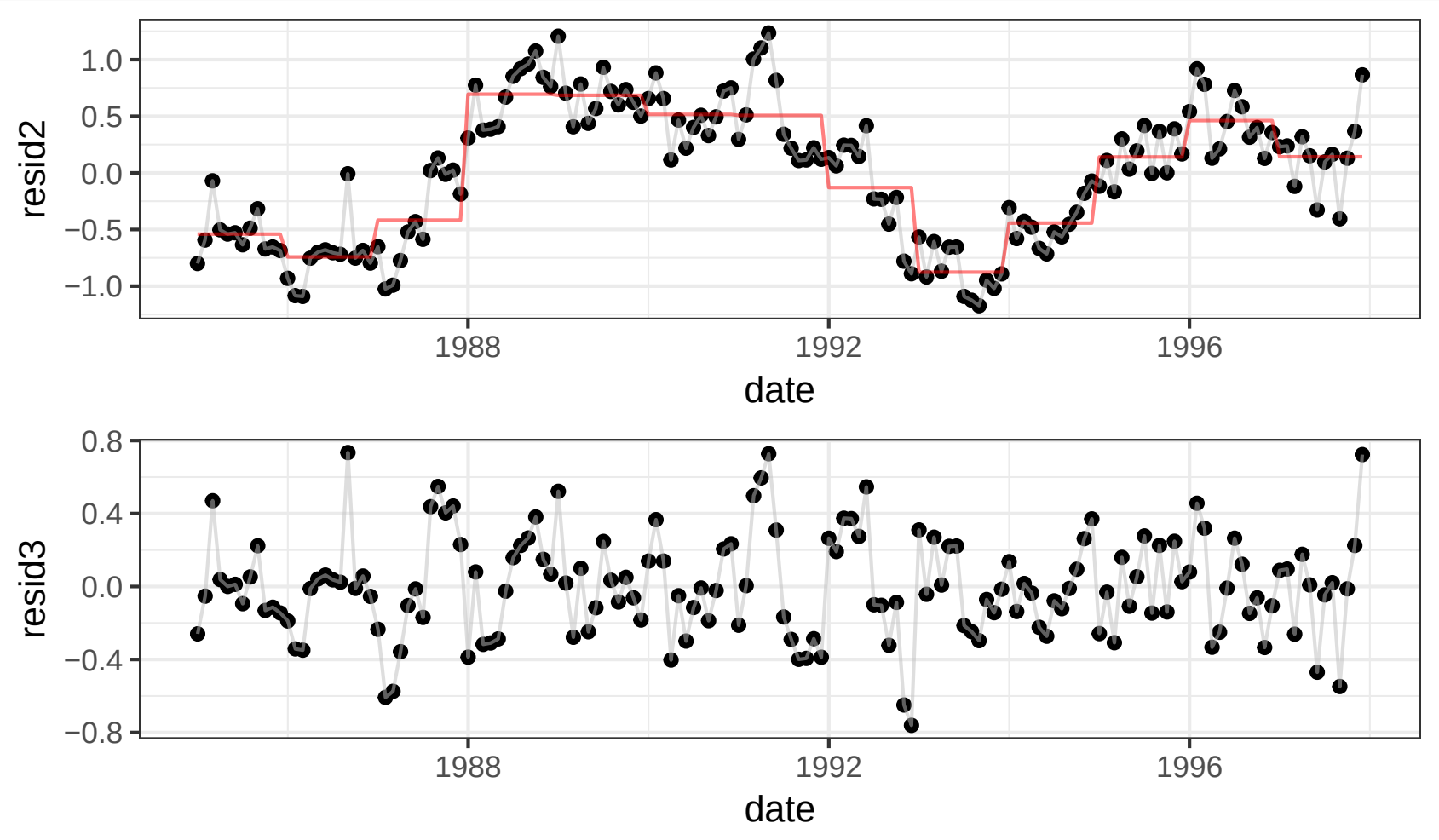


# and then and then?

Maybe there is some different effect by year ...

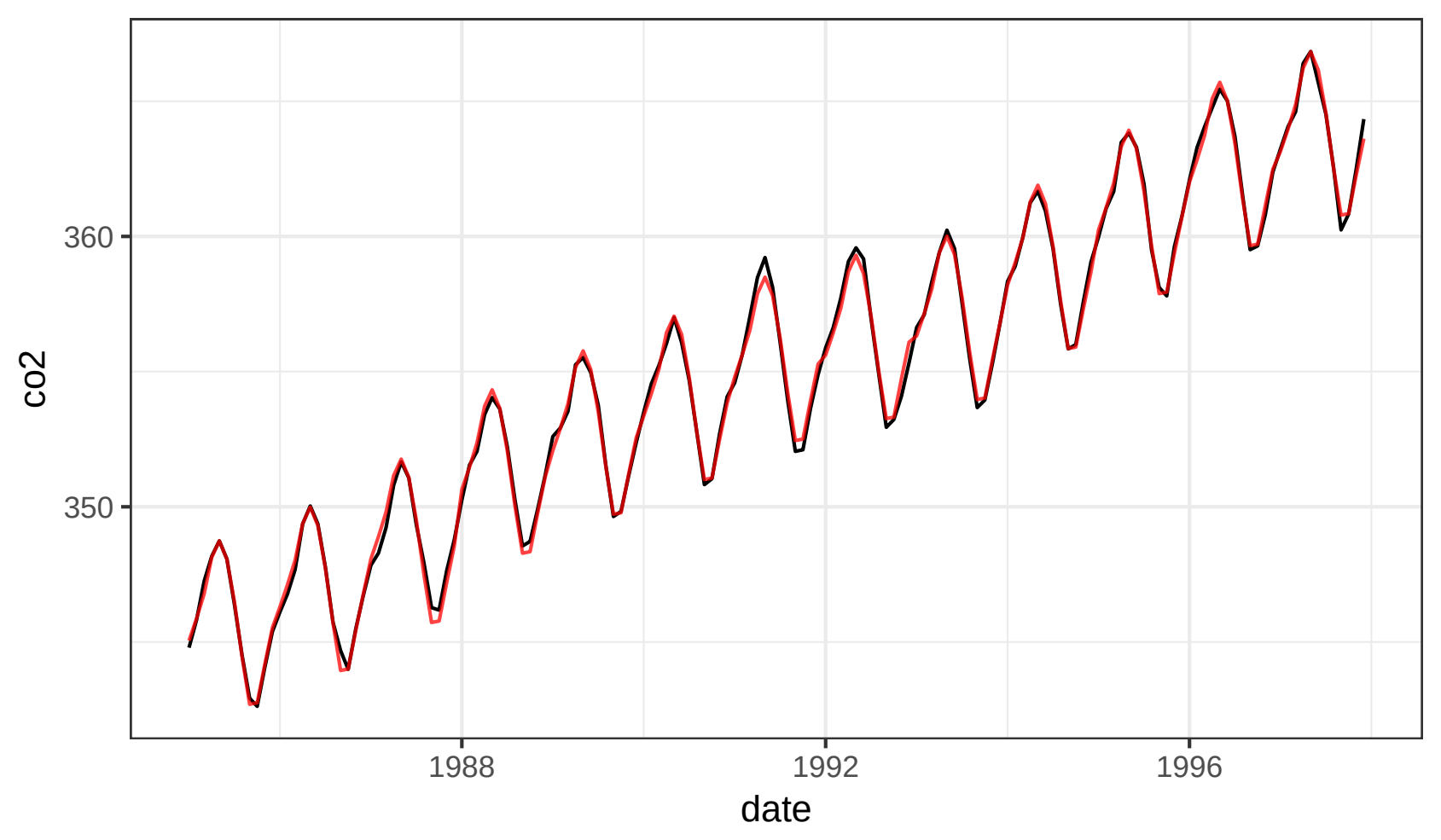
# and then and then?

Maybe there is some different effect by year ...



# Too much

```
(lm = lm(co2~date + month + as.factor(year), data=co2_df))
##
## Call:
## lm(formula = co2 ~ date + month + as.factor(year), data = co2_df)
##
## Coefficients:
##          (Intercept)                  date              monthAug
##            -2.645e+03             1.508e+00             -4.177e+00
##              monthDec              monthFeb              monthJan
##            -3.612e+00             -2.008e+00             -2.705e+00
##              monthJul              monthJun              monthMar
##            -2.035e+00             -3.251e-01             -1.227e+00
##              monthMay              monthNov              monthOct
##             4.821e-01             -4.838e+00             -6.135e+00
##              monthSep  as.factor(year)1986  as.factor(year)1987
##            -6.064e+00             -2.585e-01              9.722e-03
## as.factor(year)1988  as.factor(year)1989  as.factor(year)1990
##             1.065e+00              9.978e-01              7.726e-01
## as.factor(year)1991  as.factor(year)1992  as.factor(year)1993
##             7.067e-01              1.236e-02             -7.911e-01
## as.factor(year)1994  as.factor(year)1995  as.factor(year)1996
##            -4.146e-01              1.119e-01              3.768e-01
## as.factor(year)1997
##                    NA
```


360
350
co2
1988
1992
1996
date

# Generalized Linear Models

# Background

A generalized linear model has three key components:

1. a probability distribution (from the exponential family) that describes your response variable
2. a linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$,
3. and a link function $g$ such that $g\big(E(Y|X)\big) = \boldsymbol{\mu} = \boldsymbol{\eta}$.

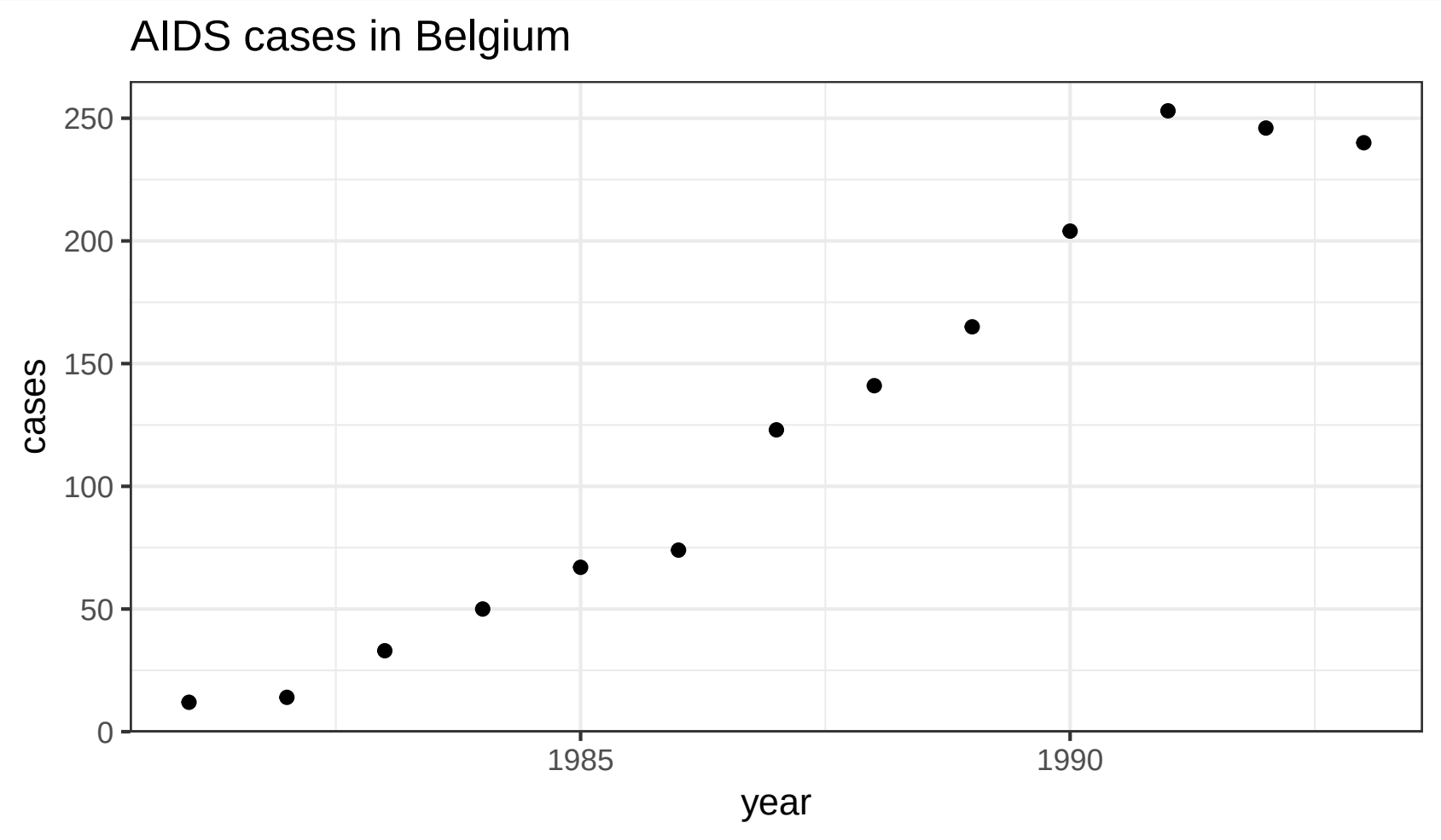
# Poisson Regression

# Model Specification

A generalized linear model for count data where we assume the outcome variable follows a poisson distribution (mean = variance).

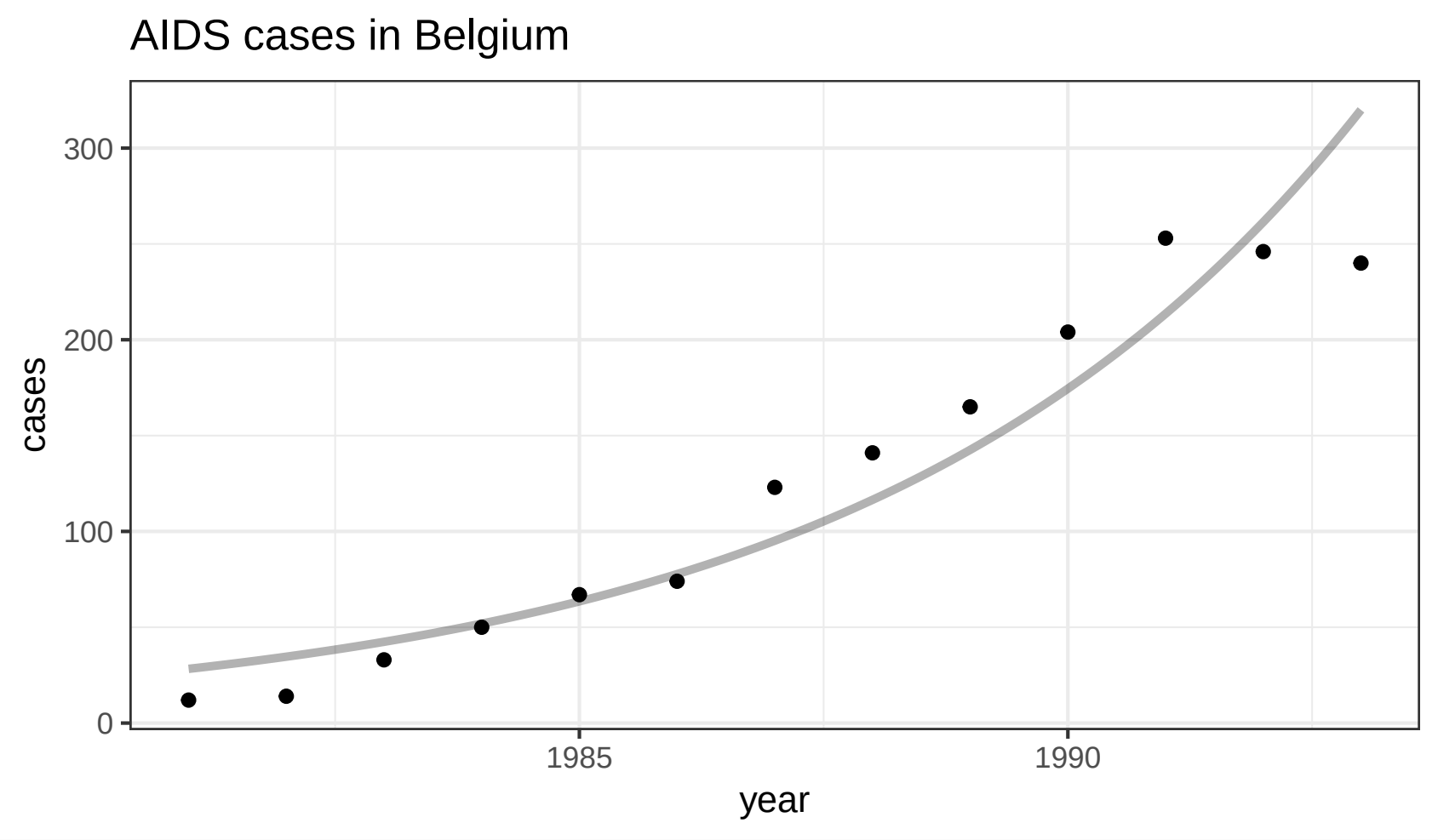
$$
\begin{aligned}
Y_i &\sim \text{Poisson}(\lambda_i) \\
\log E(\boldsymbol{Y}|\boldsymbol{X}) &= \log \boldsymbol{\lambda} = \boldsymbol{X}\boldsymbol{\beta}
\end{aligned}
$$

# Example - AIDS in Belgium



# Frequentist glm fit

```
g = glm(cases~year, data=aids, family=poisson)
pred = data_frame(year=seq(1981,1993,by=0.1))
pred$cases = predict(g, newdata=pred, type = "response")
```



# Residuals

Standard residuals:

$$r_i = Y_i - \hat{Y}_i = Y_i - \hat{\lambda}_i$$

Pearson residuals:

$$r_i = \frac{Y_i - E(Y_i|X)}{\sqrt{Var(Y_i|X)}} = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

Deviance residuals:

$$d_i = \text{sign}(y_i - \lambda_i)\sqrt{2(y_i \log(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i))}$$

# Deviance and deviance residuals

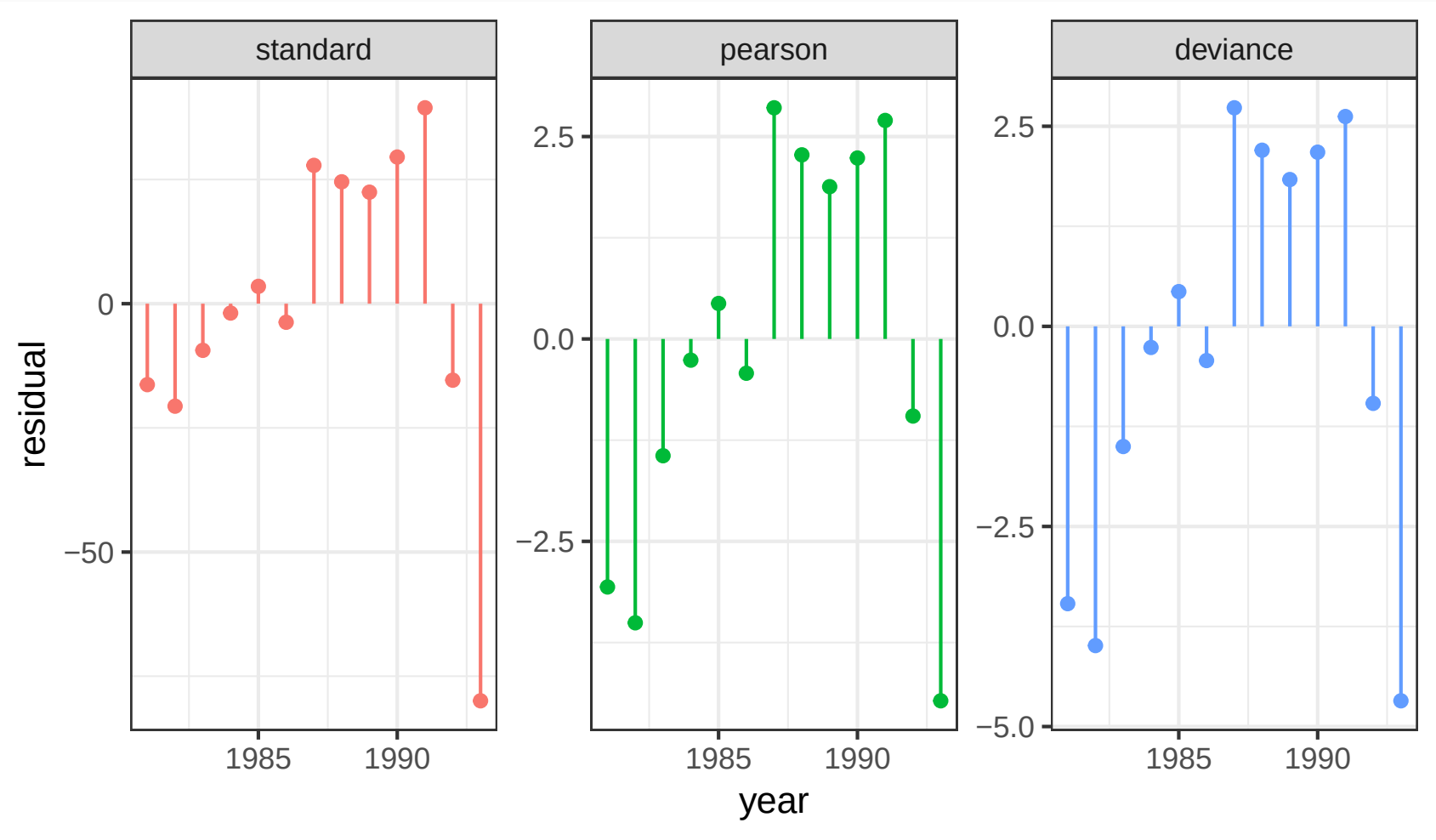
Deviance can be interpreted as the difference between your model's fit and the fit of an ideal model (where $E(\hat{Y}_i) = Y_i$).

$$D = 2(\mathcal{L}(Y|\theta_{best}) - \mathcal{L}(Y|\hat{\theta})) = \sum_{i=1}^{n} {d_i}^2$$

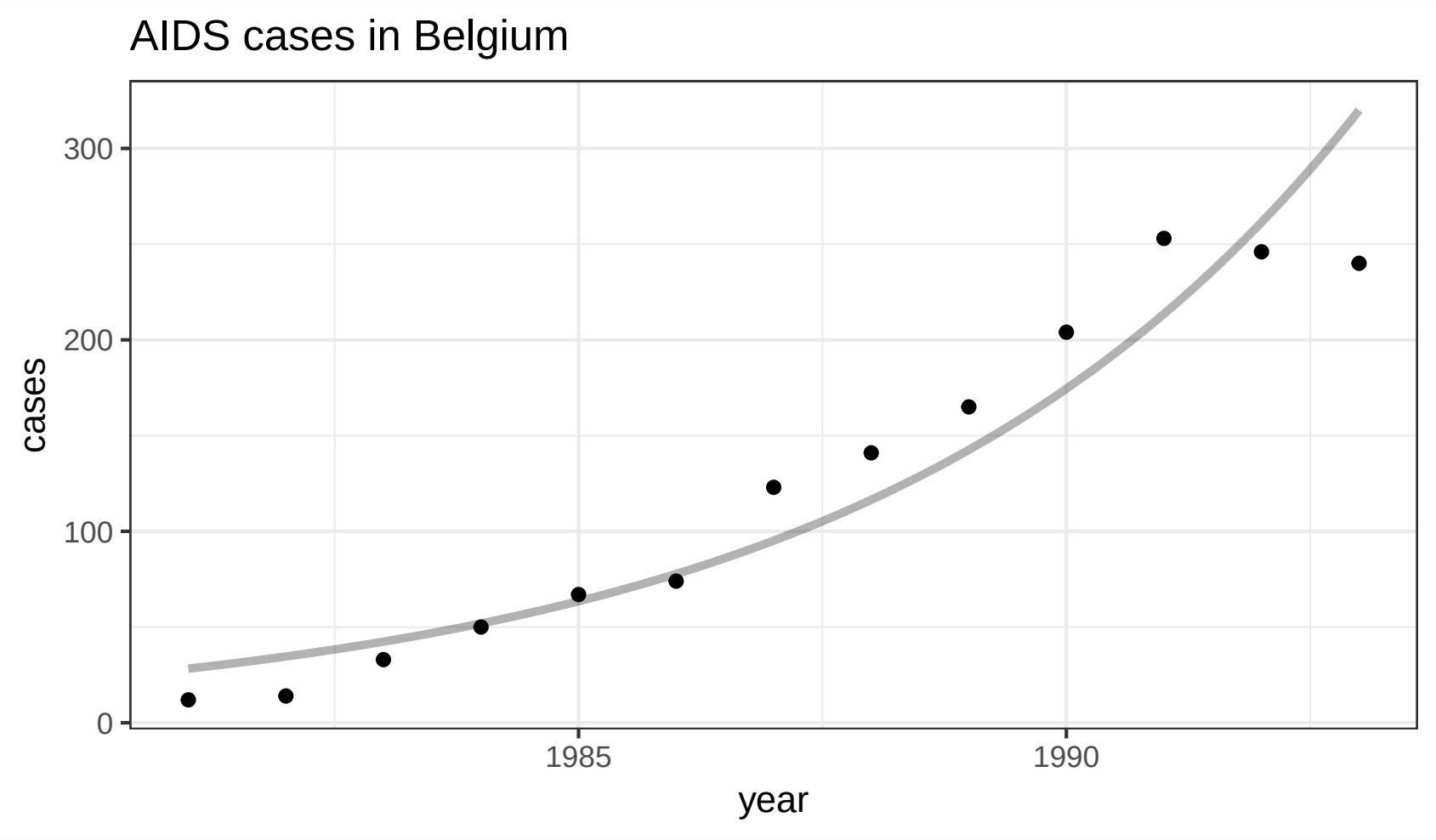
Deviance is a measure of goodness of fit in a similar way to the residual sum of squares (which is just the sum of squared standard residuals).

# Deviance residuals derivation

# Residual plots

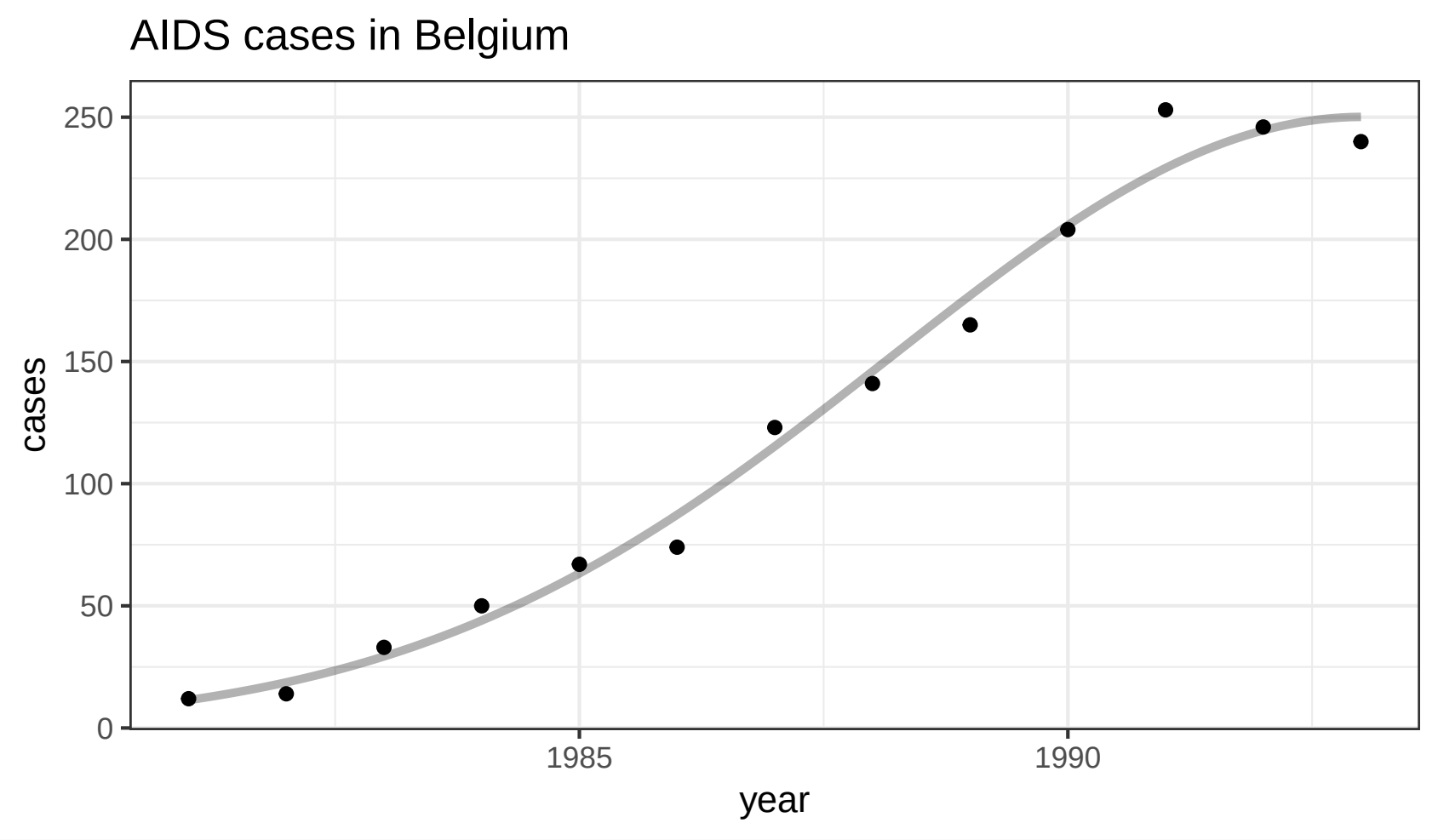


```
print(aids_fit)
```



# Quadratic fit

```
g2 = glm(cases~year+I(year^2), data=aids, family=poisson)
pred2 = data_frame(year=seq(1981,1993,by=0.1))
pred2$cases = predict(g2, newdata=pred, type = "response")
```



# Quadratic fit - residuals

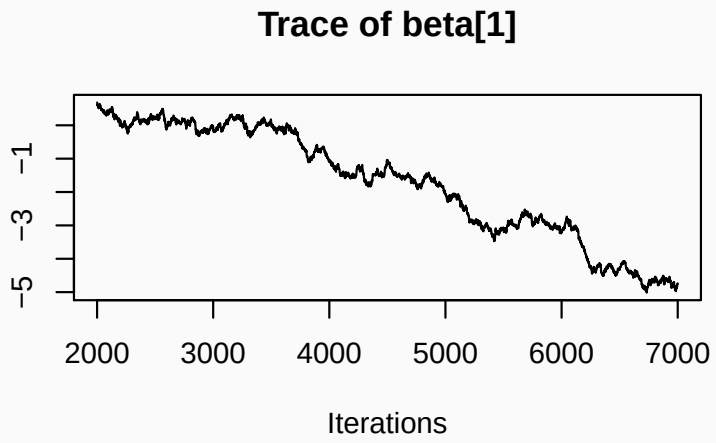
# Bayesian Poisson Regression Model

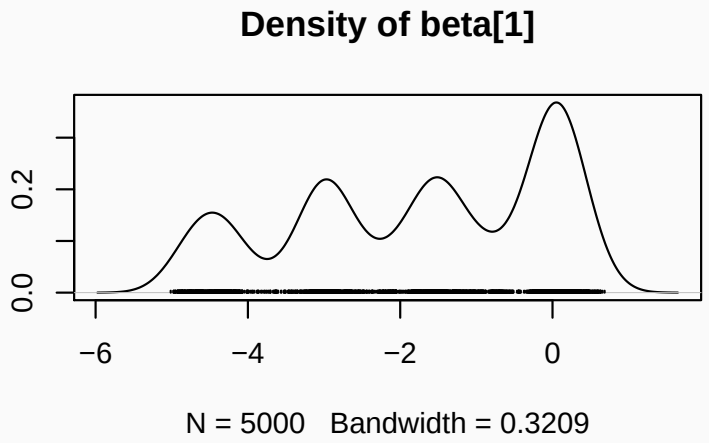
```
## model{
##   # Likelihood
##   for(i in 1:length(Y)){
##     Y[i] ~ dpois(lambda[i])
##     log(lambda[i]) <- beta[1] + beta[2]*X[i]
##
##     # In-sample prediction
##     Y_hat[i] ~ dpois(lambda[i])
##   }
##
##   # Prior for beta
##   for(j in 1:2){
##     beta[j] ~ dnorm(0,1/100)
##   }
## }
```

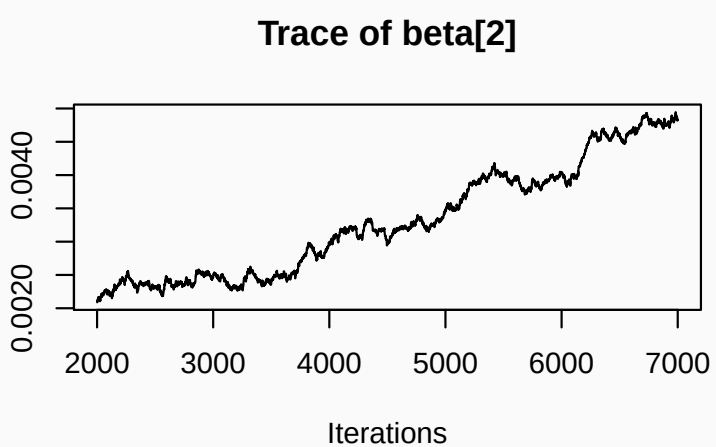
# Bayesian Model fit

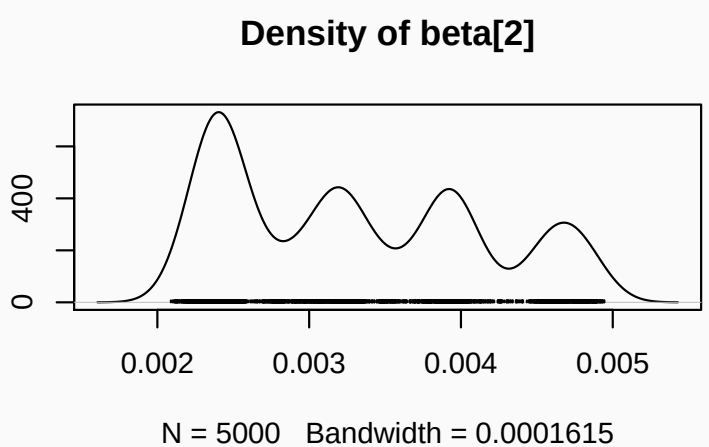
```
m = jags.model(
  textConnection(poisson_model1), quiet = TRUE,
  data = list(Y=aids$cases, X=aids$year)
)
update(m, n.iter=1000, progress.bar="none")
samp = coda.samples(
  m, variable.names=c("beta","lambda","Y_hat"),
  n.iter=5000, progress.bar="none"
)
```

# MCMC Diagnostics

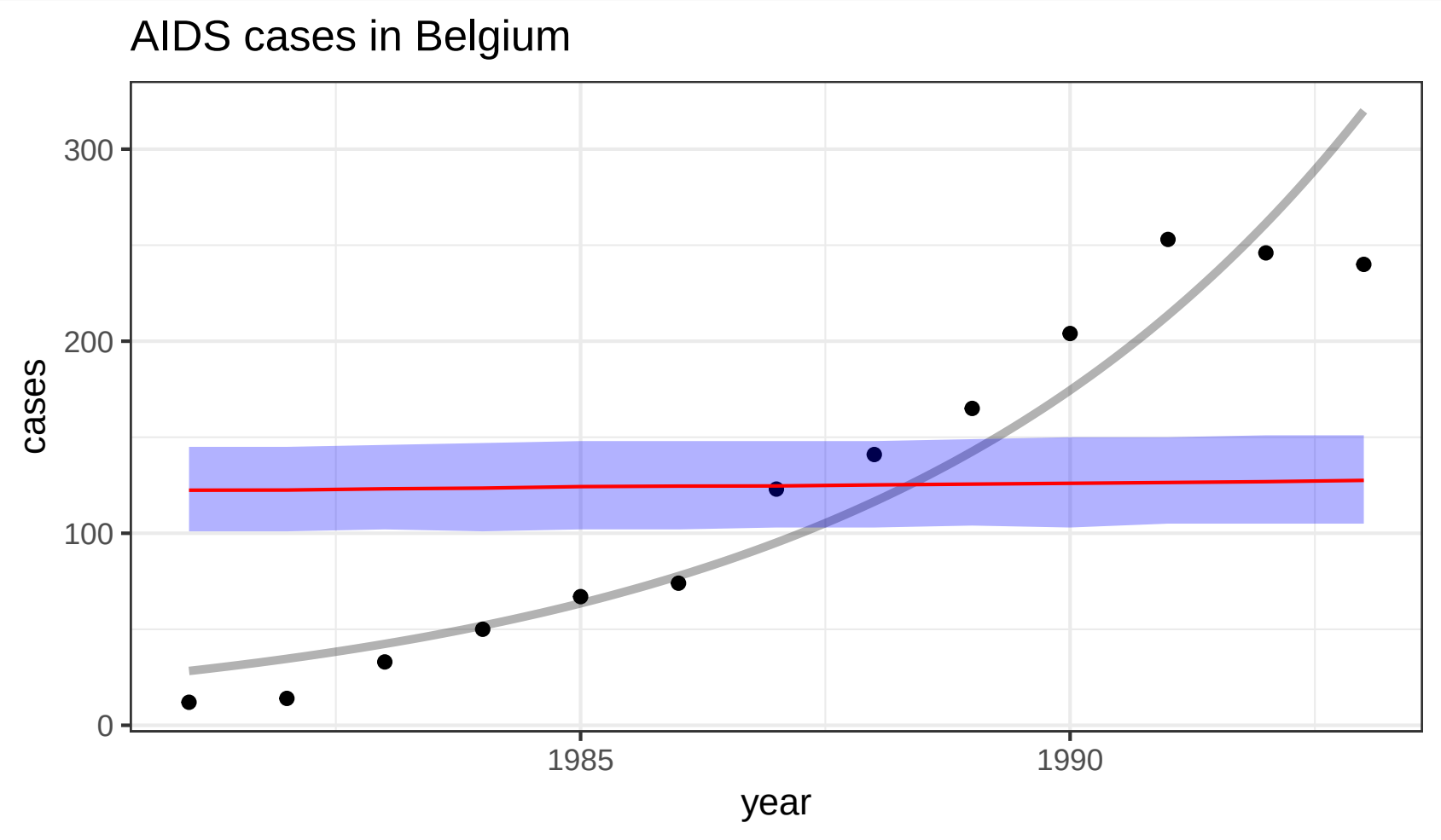


# Model fit?



# What went wrong?

# What went wrong?

```
summary(g)
##
## Call:
## glm(formula = cases ~ year, family = poisson, data = aids)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6784  -1.5013  -0.2636   2.1760   2.7306
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.971e+02  1.546e+01  -25.68   <2e-16 ***
## year         2.021e-01  7.771e-03   26.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.206  on 12  degrees of freedom
## Residual deviance:  80.686  on 11  degrees of freedom
## AIC: 166.37
##
## Number of Fisher Scoring iterations: 4
```

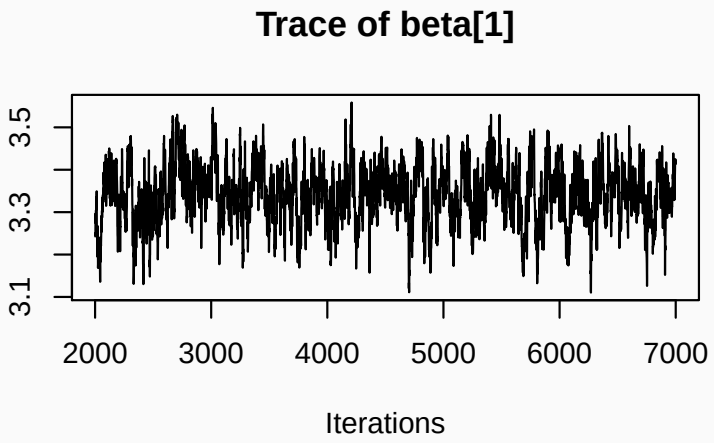
```r
summary(glm(cases~I(year-1981), data=aids, family=poisson))
```

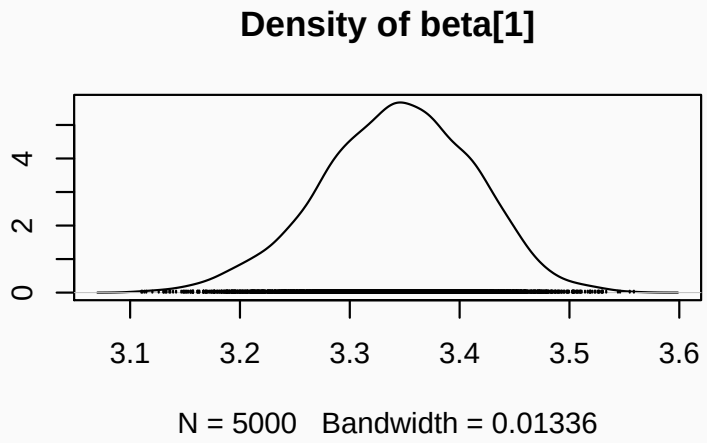
```
##
## Call:
## glm(formula = cases ~ I(year - 1981), family = poisson, data = aids)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6784  -1.5013  -0.2636   2.1760   2.7306
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.342711   0.070920   47.13   <2e-16 ***
## I(year - 1981) 0.202121   0.007771   26.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.206  on 12  degrees of freedom
## Residual deviance:  80.686  on 11  degrees of freedom
## AIC: 166.37
##
## Number of Fisher Scoring iterations: 4
```

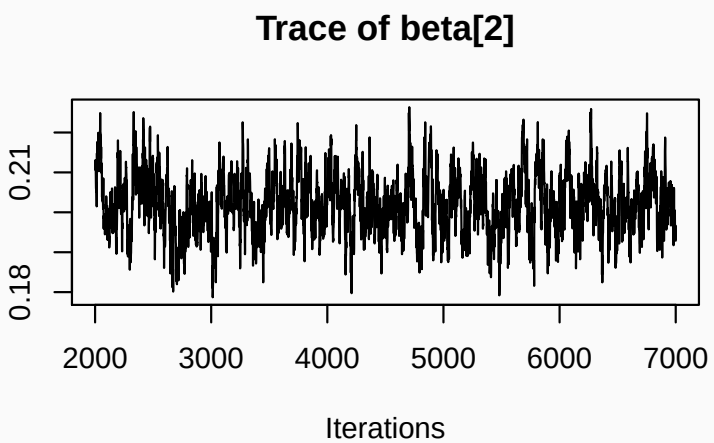
# Revising the model

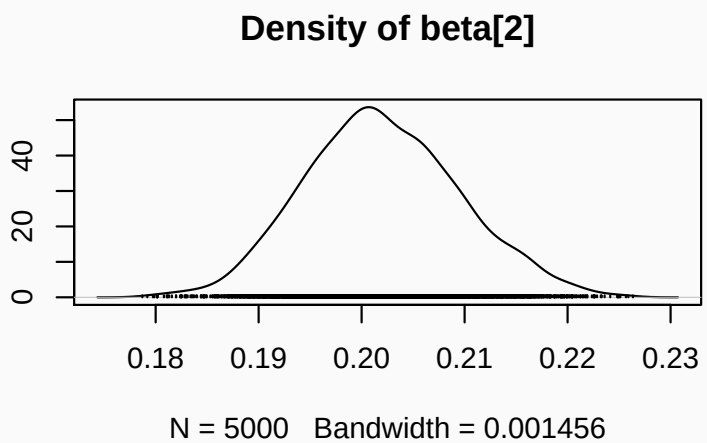
```
## model{
##   # Likelihood
##   for(i in 1:length(Y)){
##     Y[i] ~ dpois(lambda[i])
##     log(lambda[i]) <- beta[1] + beta[2]*(X[i] - 1981)
##
##     # In-sample prediction
##     Y_hat[i] ~ dpois(lambda[i])
##   }
##
##   # Prior for beta
##   for(j in 1:2){
##     beta[j] ~ dnorm(0,1/100)
##   }
## }
```

# MCMC Diagnostics

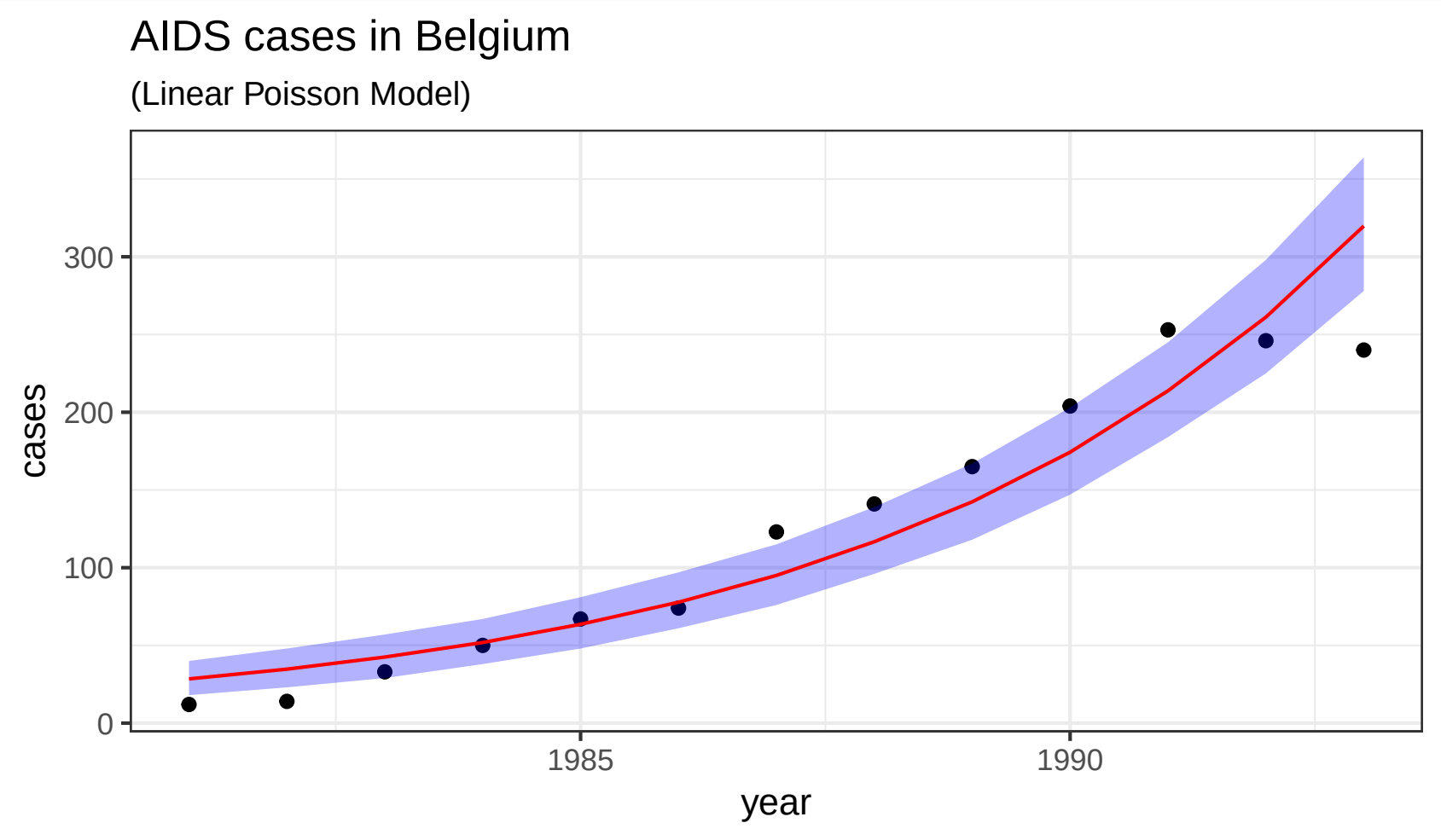


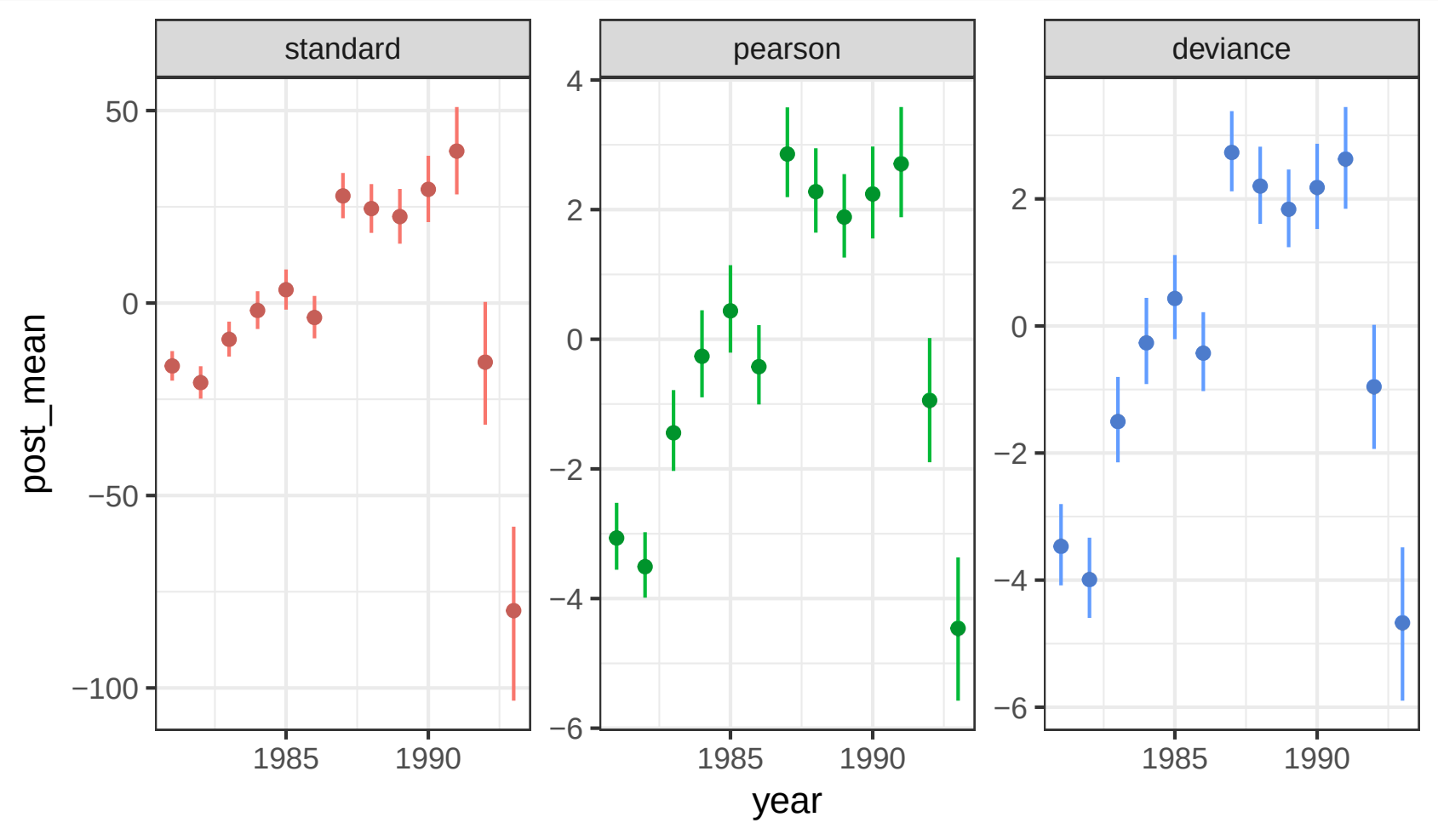



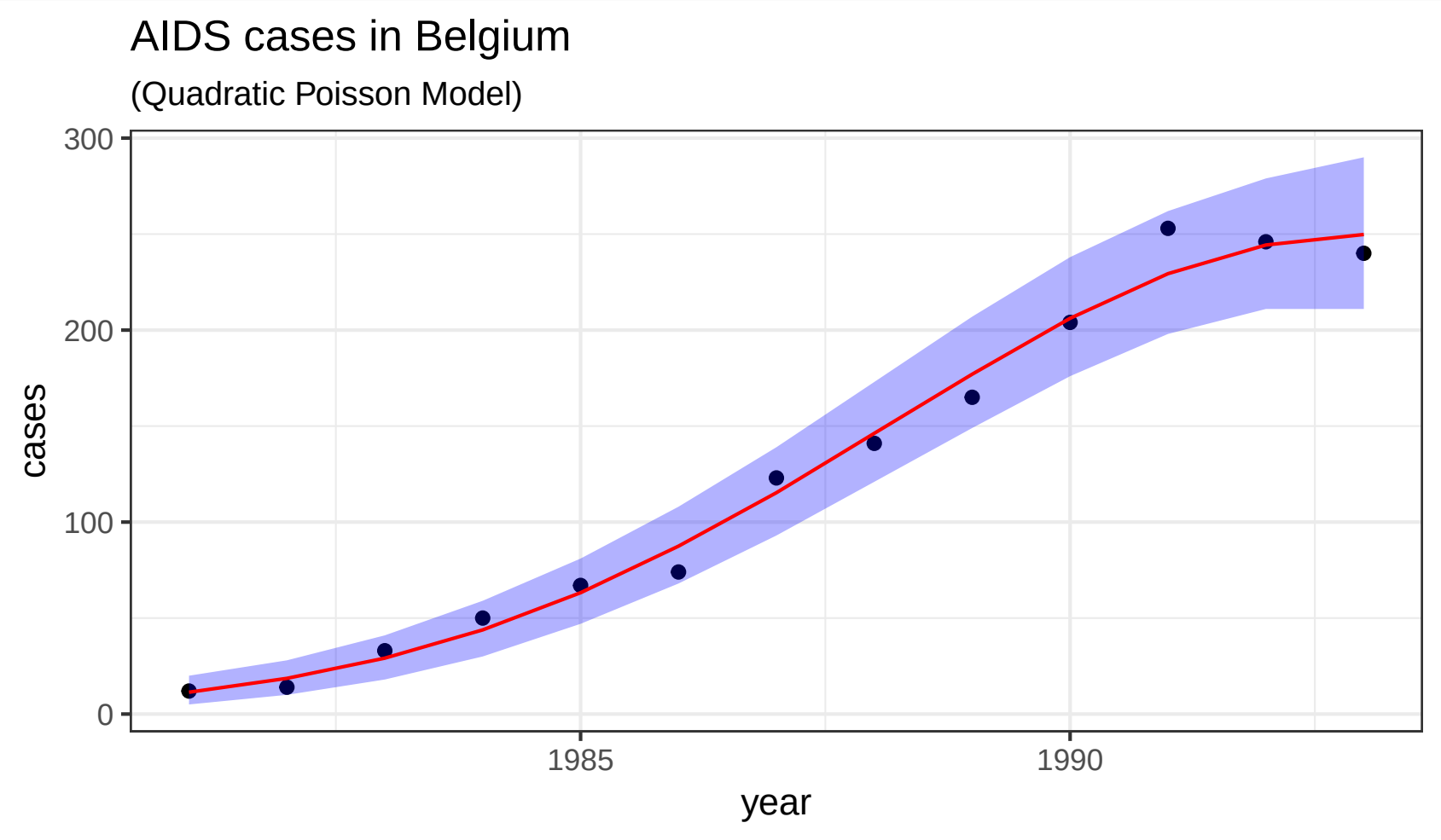
# Model fit


AIDS cases in Belgium
(Linear Poisson Model)
cases
300
200
100
0
1985
1990
year

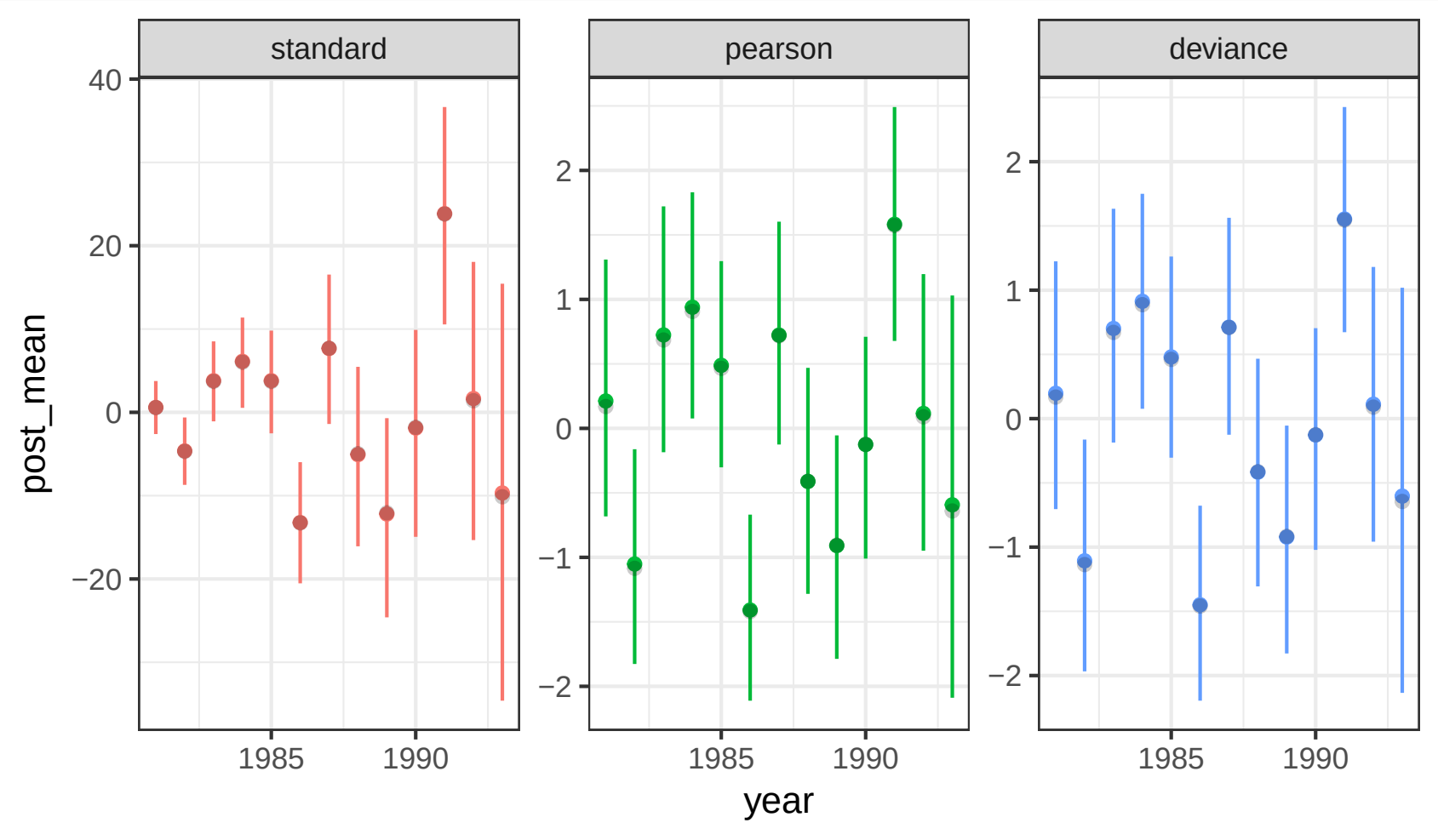
# Bayesian Residual Plots



# Model fit


AIDS cases in Belgium
(Quadratic Poisson Model)
300
200
100
0
cases
1985
1990
year

# Bayesian Residual Plots



# Negative Binomial Regression

# Overdispersion

One of the properties of the Poisson distribution is that if $X \sim \text{Pois}(\lambda)$ then $E(X) = Var(X) = \lambda$.

If we are constructing a model where we claim that our response variable $Y$ follows a Poisson distribution then we are making a very strong assumption which has implactions for both inference and prediction.

# Overdispersion

One of the properties of the Poisson distribution is that if $X \sim \text{Pois}(\lambda)$ then $E(X) = Var(X) = \lambda$.

If we are constructing a model where we claim that our response variable $Y$ follows a Poisson distribution then we are making a very strong assumption which has implactions for both inference and prediction.

```
mean(aids$cases)
## [1] 124.7692
var(aids$cases)
## [1] 8124.526
```

# Negative binomial regession

If we define

$$
\begin{aligned}
Y_i | Z_i &\sim \text{Pois}(\lambda_i \, Z_i) \\
Z_i &\sim \text{Gamma}(\theta_i, \, \theta_i)
\end{aligned}
$$

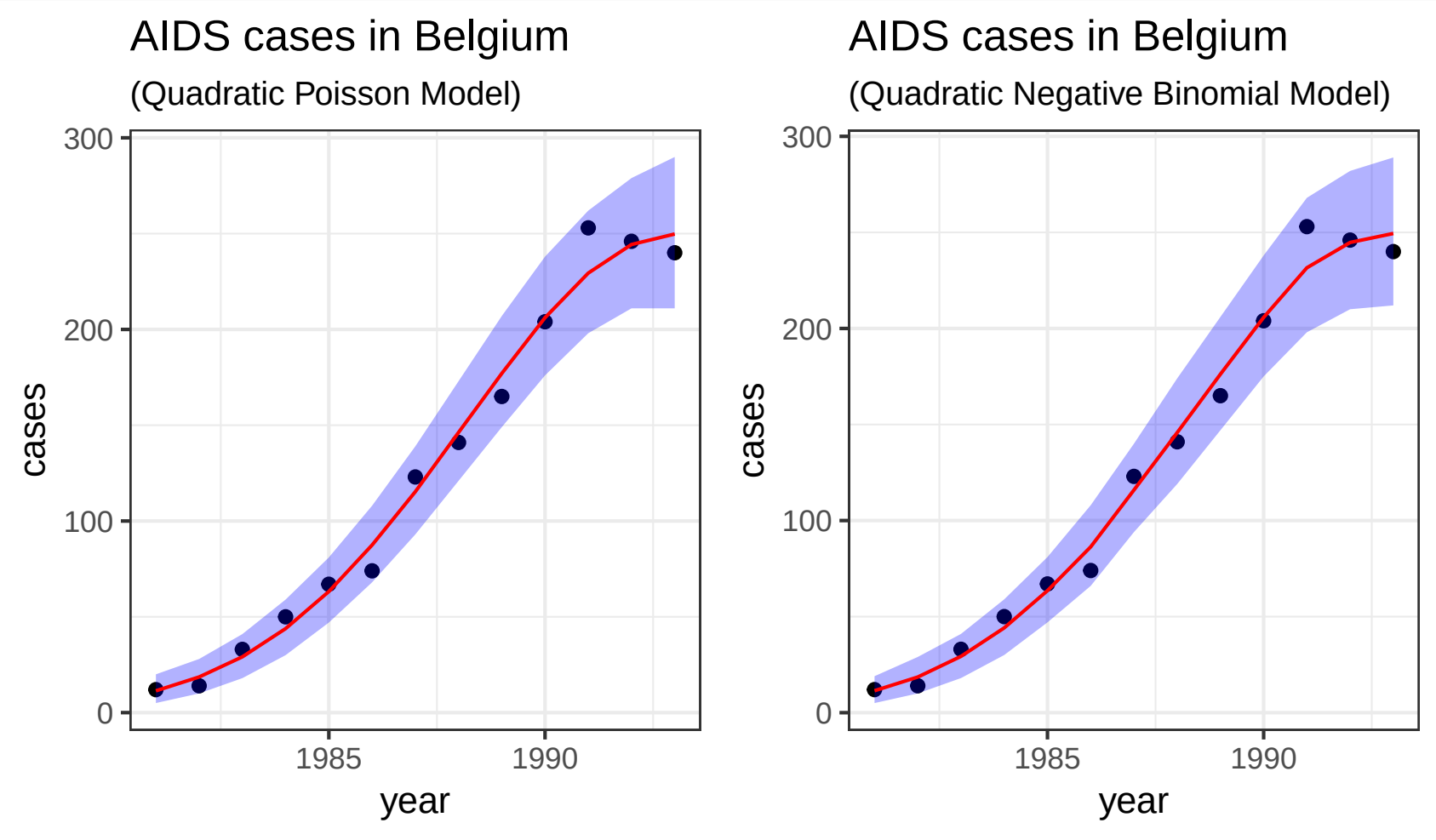
then the marginal distribution of $Y_i$ will be negative binomial with,

$$
\begin{aligned}
E(Y_i) &= \lambda_i \\
Var(Y_i) &= \lambda_i + \lambda_i^2 / \theta_i
\end{aligned}
$$

# Model

```
## model{
##   for(i in 1:length(Y))
##   {
##     Z[i] ~ dgamma(theta, theta)
##     log(lambda[i]) <- beta[1] + beta[2]*(X[i] - 1981) + beta[3]*(X[i] - 1
##
##     lambda_Z[i] <- Z[i]*lambda[i]
##
##     Y[i] ~ dpois(lambda_Z[i])
##     Y_hat[i] ~ dpois(lambda_Z[i])
##   }
##
##   for(j in 1:3){
##     beta[j] ~ dnorm(0, 1/100)
##   }
##
##   log_theta ~ dnorm(0, 1/100)
##   theta <- exp(log_theta)
## }
```

# Negative Binomial Model fit



# Bayesian Residual Plots