intro

Ben Bolker

04 Sep 2018

Logistics

- contact info, e-mail policies
- textbook
- assignments & grading
- policies: group work, take-home exams, etc.

Scope

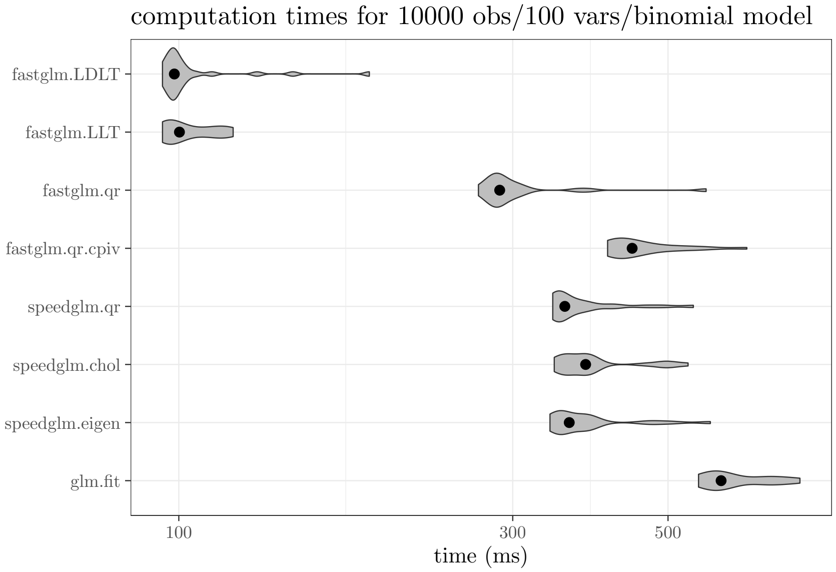
- Topics
  - core:
    - linear models: design matrices, contrasts, etc.
    - core GLMs: binary (logistic/probit), binomial, Poisson regression
    - weird GLMs and further topics: complete separation, overdispersion, Gamma models, non-standard links, use of offsets
    - more weird GLMs: ordinal, negative binomial, zero-inflated
    - GL mixed Ms: longitudinal / hierarchical / multilevel models
    - Bayesian methods
  - "extraneous"
    - data wrangling, visualization, and reproducible research: R, `ggplot`, tidyverse, Rmarkdown
    - data visualization; graphical approaches to diagnostics and model interpretation
    - best practices/ethics for data analysis
- Procedures
  - data exploration
  - model fitting (estimation)
  - graphical and numerical diagnostics
  - inference
    (Wald, likelihood, bootstrapping, AIC, . . .)
  - verbal and graphical presentation/interpretation of results

What is a GLM?

- handles any linear model
- *link function* specifies nonlinearity between linear predictor and response
- response distribution from the *exponential family* (Gaussian, binomial, Poisson, Gamma, . . .)

Why GLMs?

- robust
- fast
- sensible, flexible statistical models
- "sweet spot" in generality and power

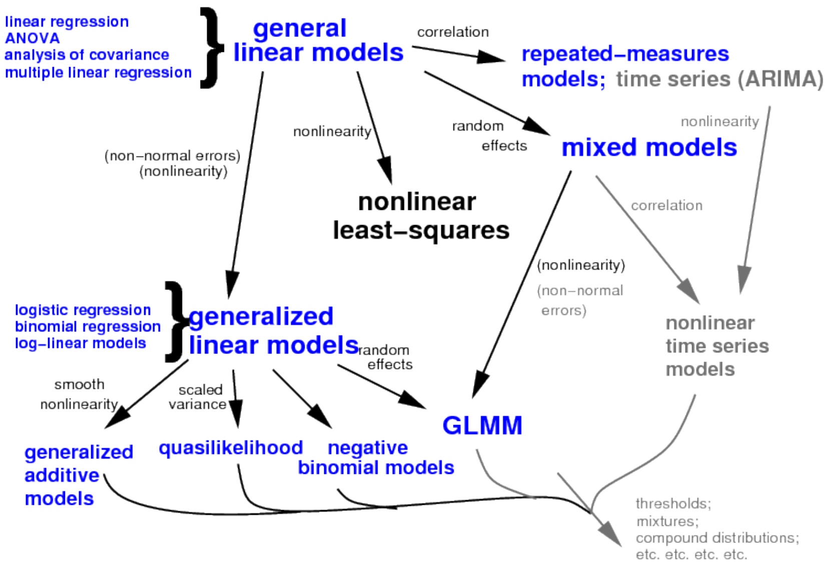


Example

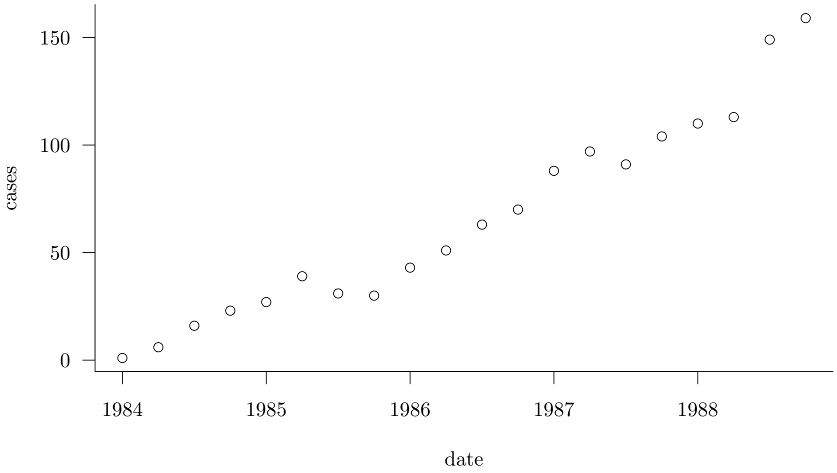
Using data on AIDS diagnoses from Australia (Dobson and Barnett p. 69). Read in data and inspect it:

```
aids <- read.csv("../data/aids.csv")
head(aids)          ## beginning of data
summary(aids)       ## min/mean/max etc.
skimr::skim(aids)  ## fancier
## construct useful date/index variables
aids <- transform(aids,
                  date=year+(quarter-1)/4,
                  index=seq(nrow(aids)))
```

Some basic pictures: base graphics

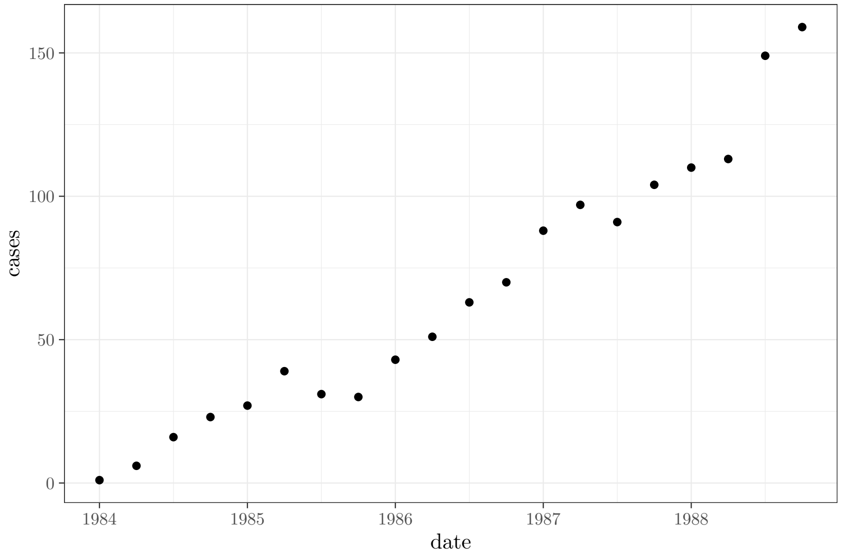


```
with(aids,plot(date,cases))
```



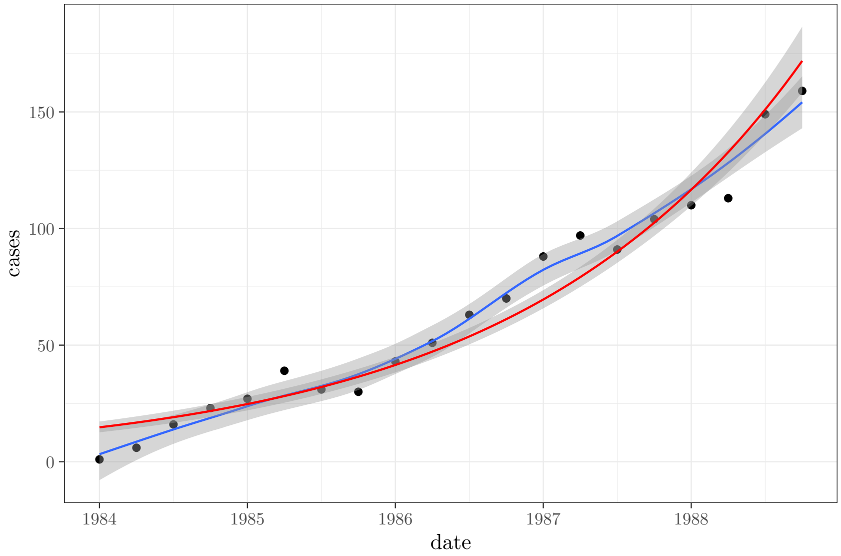
or with ggplot2

```
library(ggplot2)
theme_set(theme_bw())  ## get rid of grey background
 ## simple X/Y scatterplot
p0 <- (ggplot(aids,aes(x=date,y=cases))
    + geom_point()     ## add points
)
print(p0)
```



Now pictures with nonparametric and GLM fits superimposed:

```
(p0
    + geom_smooth()  ## nonparametric
    +  geom_smooth(method="glm",
                   method.args=list(family=poisson),
                   colour="red") ## GLM fit
)
```

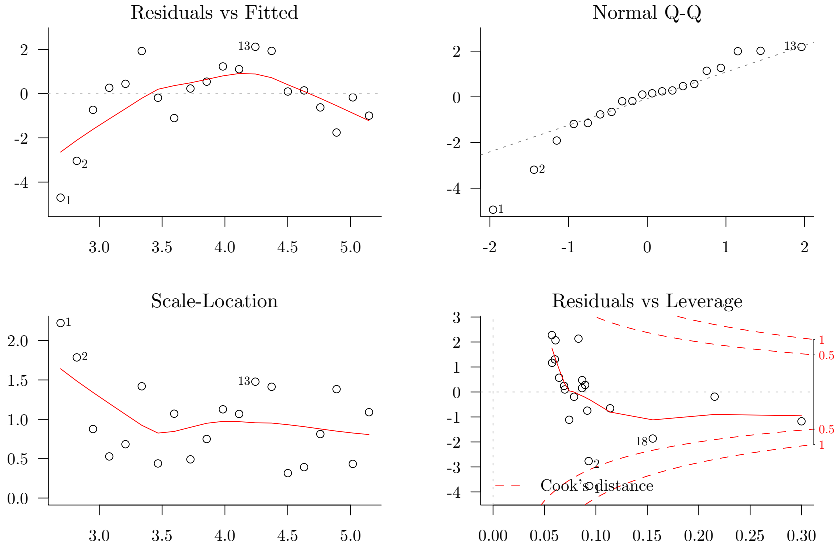


Fit a model using `glm()`:

```
g1 <- glm(cases~date, data=aids, family=poisson)
```

Diagnostic plots:

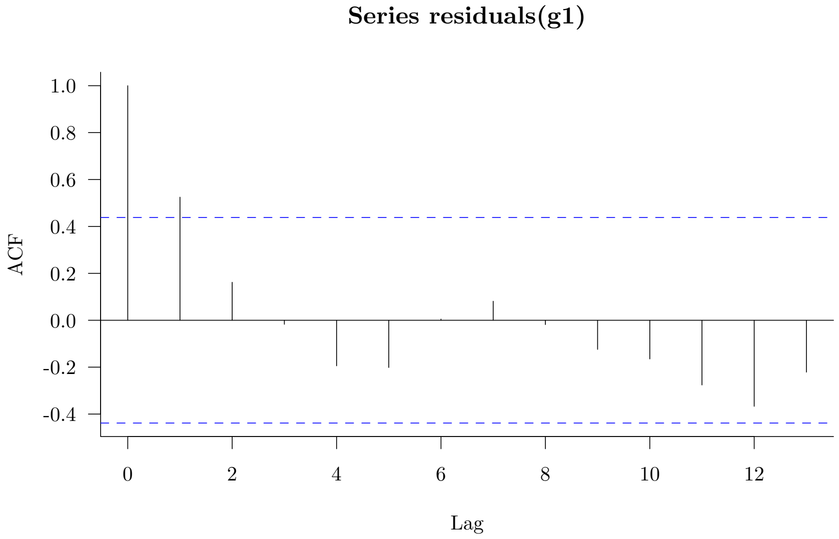
```
## set 2x2 grid of plots, tweak margins, label orientation
op <- par(mfrow=c(2,2),mar=c(3,3,2,2),
        las=1,bty="l")
plot(g1)  ## plot standard diagnostics
```



```
par(op)  ## restore parameter settings
```

Check for temporal autocorrelation:

```
acf(residuals(g1))
```



We have some problems. Will a quadratic fit help?

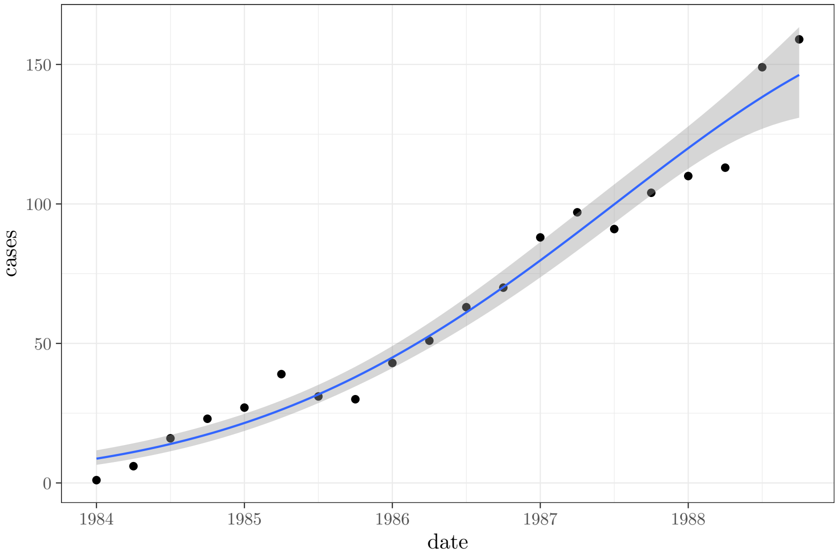
```
## poly(.,2) sets up a degree-2 (quadratic) polynomial
g2 <- glm(cases~poly(date,2),aids,family=poisson)
summary(g2)  ## quadratic term significantly negative

##
```

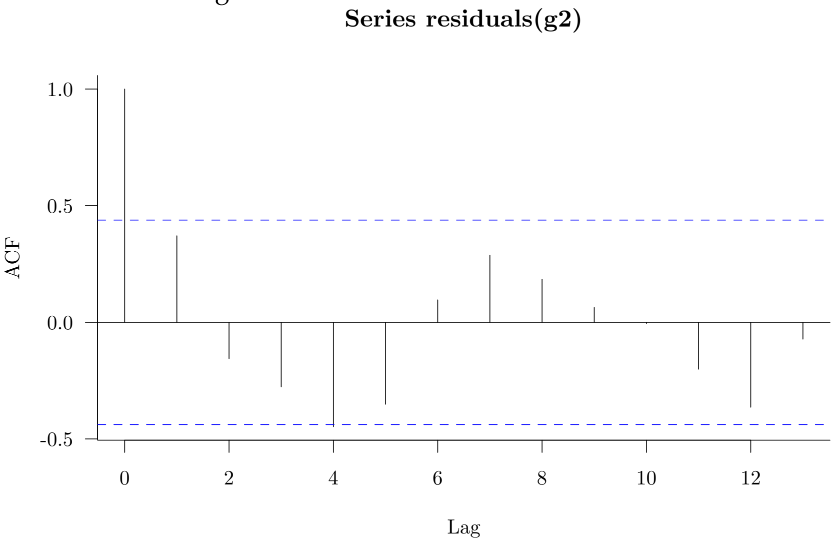
```
## Call:
## glm(formula = cases ~ poly(date, 2), family = poisson, data = aids)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3290  -0.9071  -0.0761   0.8985   2.3209
##
## Coefficients:
##                 Estimate Std. Error z value
## (Intercept)      3.86859    0.03887  99.528
## poly(date, 2)1   3.82934    0.19545  19.592
## poly(date, 2)2  -0.68335    0.15315  -4.462
##                 Pr(>|z|)
## (Intercept)      < 2e-16 ***
## poly(date, 2)1   < 2e-16 ***
## poly(date, 2)2  8.12e-06 ***
## ---
## Signif. codes:
##   0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1  ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 677.264  on 19  degrees of freedom
## Residual deviance:  31.992  on 17  degrees of freedom
## AIC: 150.29
##
## Number of Fisher Scoring iterations: 4
```

A picture of the same model fit:

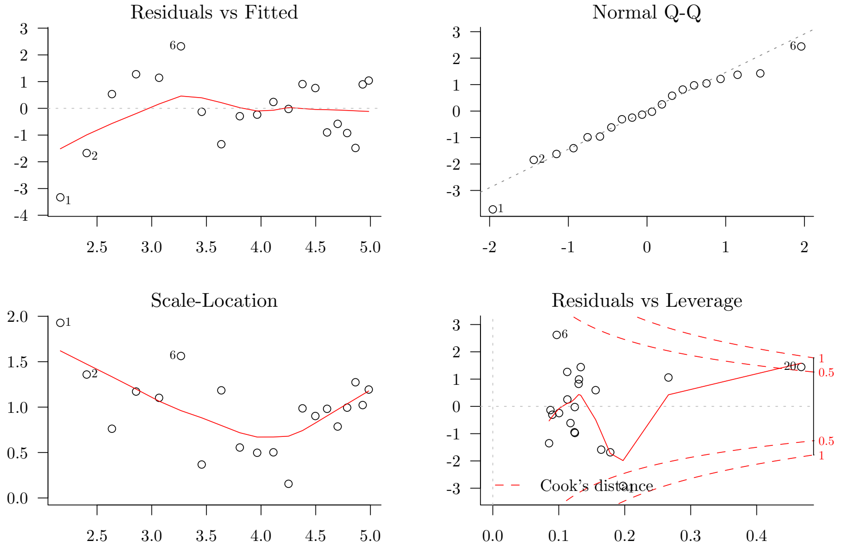
```
(p0
    +geom_smooth(method="glm",
                 formula=y~poly(x,2),
                 method.args=list(family=poisson))
)
```



Looks like the diagnostics and autocorrelation are better now . . .



```
op <- par(mfrow=c(2,2),mar=c(3,3,2,2),
          las=1,bty="l")  ## tweak params as before
plot(g2)
```



```
par(op)  ## restore parameter settings
```

Power-law model

Despite stating that "[i]n the early phase of the epidemic, the numbers of cases seemed to be increasing exponentially", Dobson and Barnett (2008) suggest fitting a power-law model of the form $Y \sim \text{Poisson}(\lambda = t^\theta)$ to the data instead:

```
g3 <- glm(cases~log(index),data=aids,family=poisson)
```

This fits pretty well, in fact much better than even the Gaussian (quadratic-exponential) model (not shown ...).

```
##             Estimate Std. Error z value
## (Intercept)   0.9960     0.1697    5.87
## log(index)    1.3266     0.0646   20.53
##             Pr(>|z|)
## (Intercept)  4.4e-09 ***
## log(index)   < 2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

- The intercept is near 1; did we already know that 1984 was the origination year of AIDS in Australia (in which case AIDS(1)=1)?
- The power law model is AIDS$(t) \propto t^{1.33}$, with 95% confidence intervals on the exponent of $\{1.2, 1.46\}$ — what does that mean biologically/epidemiologically?

This turns out, like almost every problem, to be interesting and a bit challenging when you look at it carefully (see Andrew Gelman on "god is in every leaf of every tree" - but also consider Tukey "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" or Grenfell "don't overegg the pudding"

References

Dobson, Annette J., and Adrian Barnett. 2008. *An Introduction to Generalized Linear Models, Third Edition*. 3rd ed. Chapman; Hall/CRC.