

STAT 4/6Co3: notes, class 1 (introduction)

Ben Bolker

September 6, 2013



Licensed under the Creative Commons attribution-noncommercial license

(<http://creativecommons.org/licenses/by-nc/3.0/>). Please share & remix

noncommercially, mentioning its origin.

Version: 2013-09-06 12:24:21

Logistics

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

Scope

- Topics
 - computation, visualization, and reproducible research: R, ggplot, knitr
 - 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
... general MLE
 - ordinal/nominal data (survival analysis?)
 - GL mixed Ms: longitudinal / hierarchical / multilevel models
- Procedures
 - exploratory graphics/diagnostics
 - model fitting (estimation)
 - graphical and numerical diagnostics
 - inference
(Wald, score, likelihood ratio test, bootstrapping, AIC, ...)
 - verbal and graphical presentation/interpretation of results

Example

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

```
aids <- read.csv("aids.csv")
head(aids)

##   year quarter cases
## 1 1984         1     1
## 2 1984         2     6
## 3 1984         3    16
## 4 1984         4    23
## 5 1985         1    27
## 6 1985         2    39

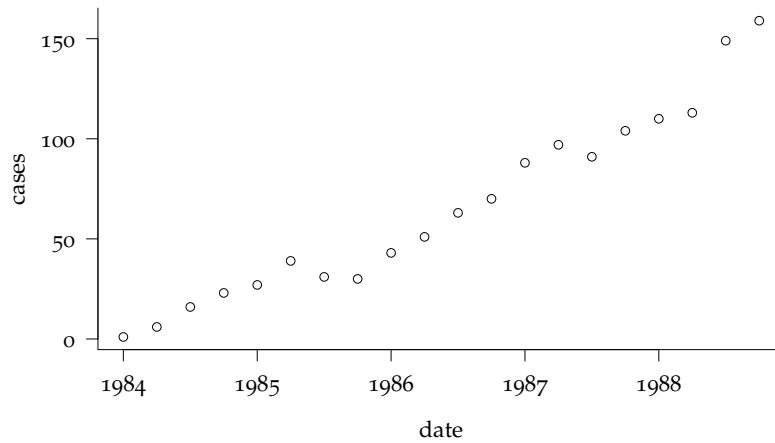
summary(aids)

##      year      quarter      cases
## Min.   :1984   Min.   :1.00   Min.   :  1.0
## 1st Qu.:1985   1st Qu.:1.75   1st Qu.: 29.2
## Median :1986   Median :2.50   Median : 57.0
## Mean   :1986   Mean   :2.50   Mean   : 65.5
## 3rd Qu.:1987   3rd Qu.:3.25   3rd Qu.: 98.8
## Max.   :1988   Max.   :4.00   Max.   :159.0

## 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))
```

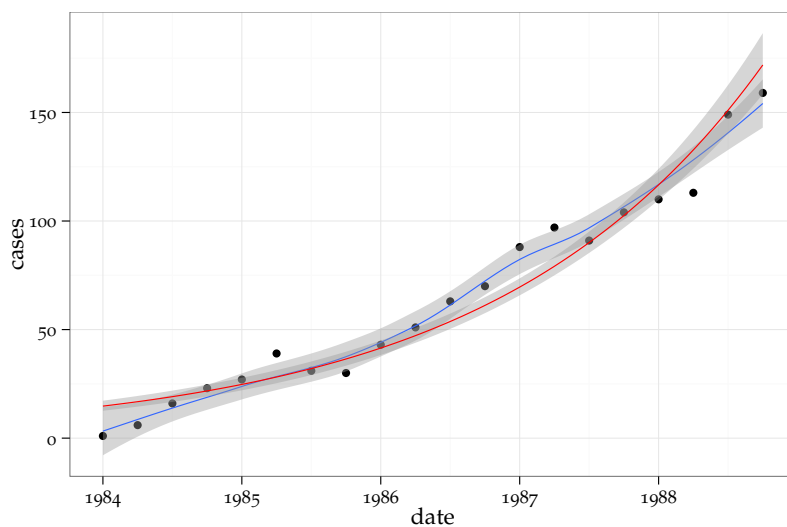


or with ggplot2

```
library(ggplot2)
theme_set(theme_bw()) ## get rid of grey background
p0 <- qplot(date,cases,data=aids) ## simple X/Y scatterplot
```

Now pictures with nonparametric and GLM fits superimposed:

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

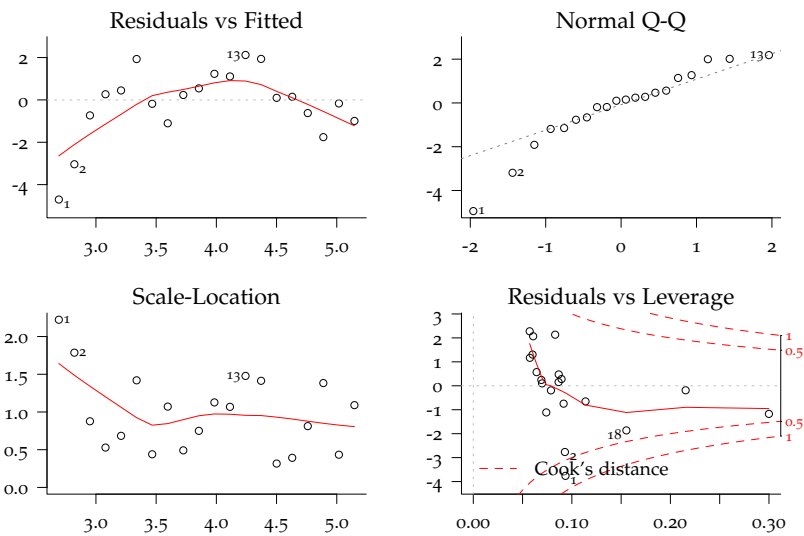


Fit a model:

```
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
```

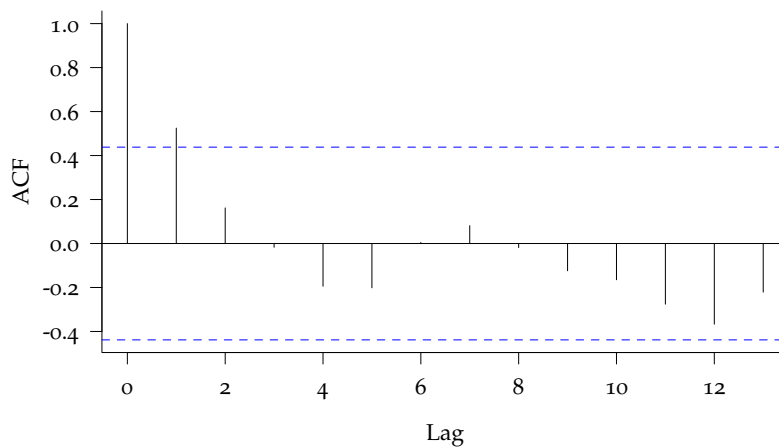


```
par(op) ## restore parameter settings
```

Check for autocorrelation as well:

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

Series residuals(g1)



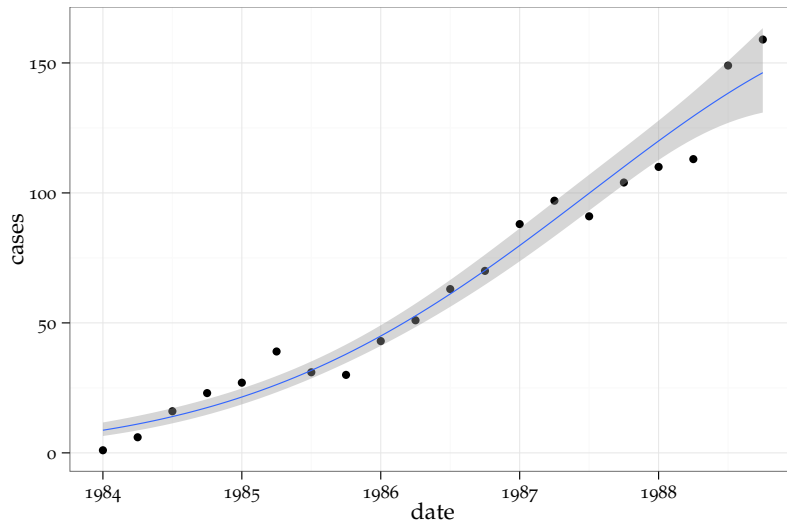
We have some problems. Will a quadratic fit help?

```
## 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
## -3.329  -0.907  -0.076   0.898   2.321
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.8686     0.0389   99.53 < 2e-16 ***
## poly(date, 2)1    3.8293     0.1955   19.59 < 2e-16 ***
## poly(date, 2)2   -0.6834     0.1532   -4.46 8.1e-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: 150.3
##
## Number of Fisher Scoring iterations: 4
```

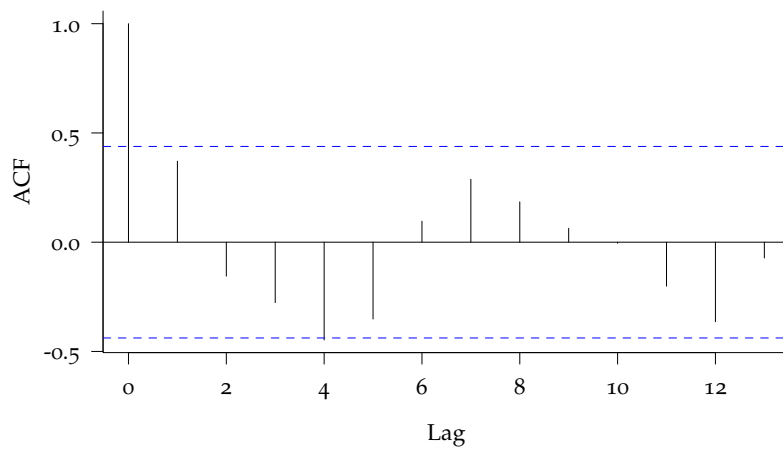
A picture of the same model fit:

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

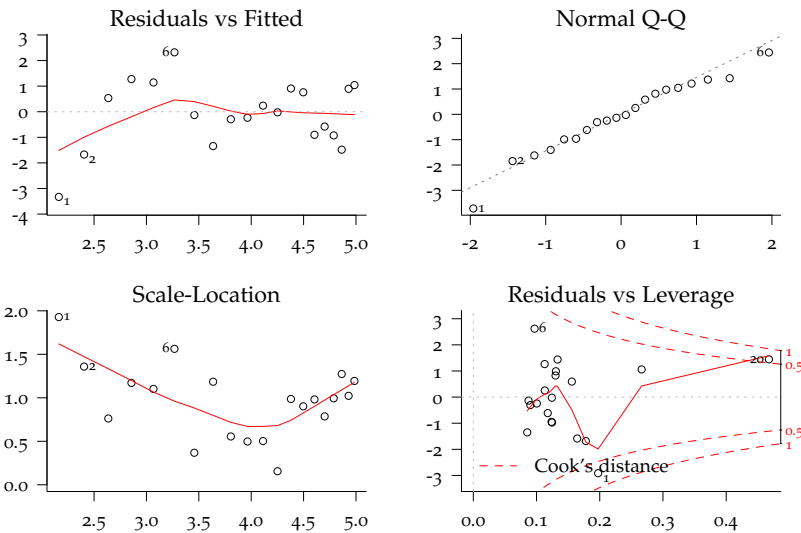


Looks like the diagnostics and autocorrelation are better now ...

Series residuals(g2)



```
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 suggest fitting a power-law model of the form $Y \sim \text{Poisson}(\lambda = t^\theta)$ to the data instead:

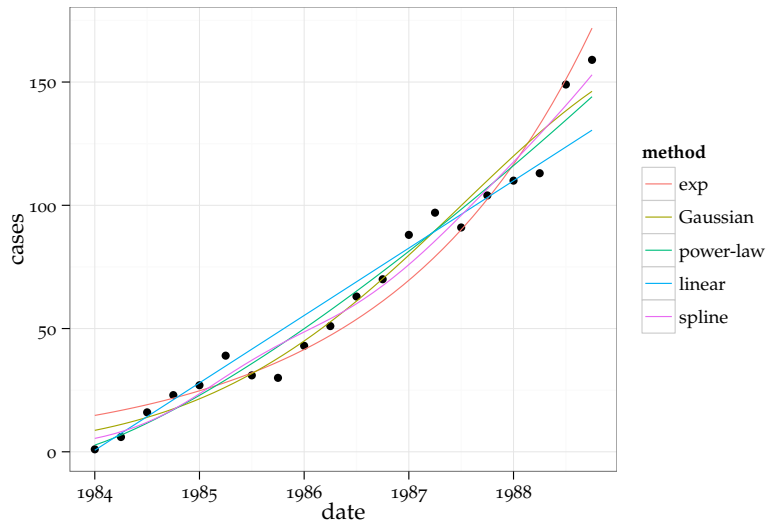
```
g3 <- glm(cases~log(index), data=aids, family=poisson)
```

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

```
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.9960    0.1697   5.87 4.4e-09 ***
## log(index)   1.3266    0.0646  20.53 < 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 $\text{AIDS}(1)=1$)?
- The power law model is $\text{AIDS}(t) \propto t^{1.33}$, with 95% confidence intervals on the exponent of $\{1.2, 1.46\}$ — what does that mean biologically/epidemiologically?

There are even more models we could try:



(confidence intervals suppressed for clarity).

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

- [1] Dobson, A. J. and A. Barnett (2008). *An Introduction to Generalized Linear Models* (3d ed.). Chapman and Hall/CRC.