

Review of last day

Poisson, log-linear regression

Interpretation of  $\beta$ 's

$$\mu = e^{\text{offset}} \times e^{\beta_0 + \beta_1 X_1 + \dots + \beta_j X_j}$$

 $\mu_i, i=1, 2$  at different  $x$ -values  
(same offset)

$$\mu_2 = \mu_1 \times \prod_{j=1}^k e^{\beta_j \Delta x_j}$$

when

$$\Delta x_j = x_{2j} - x_{1j}$$

 $\Rightarrow e^{\beta_j}$  (if other  $x$ 's held fixed)

is ~~the~~ the proportional increase in  $\mu_2$  relative to  $\mu_1$  for  $\Delta x = 1$

$\rightarrow$  so examine  $e^{\hat{\beta}_j}$

$e$  transform Cont Int.

If  $\hat{\beta}_j$  is small,

$$100 \times \hat{\beta}_j \approx \% \text{ increase}$$

Model checking

Pearson residual

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \approx N(0, 1)$$

Deviance residual  $r_i^D = \sqrt{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)^2}$   
 $\times \text{sign}(y_i - \hat{\mu}_i)$

motivation is

Deviance (analog of  $\sum (\text{Residuals})^2$ )

$$= \sum_{i=1}^n (r_i^D)^2 \quad \downarrow \quad \sum (y_i - \hat{y}_i)^2$$

$$\text{Deviance} = -2(\log \text{lik}(\hat{\beta}) - \log \text{lik}(\text{sat model}))$$

convenient benchmark

Default residual

plots provided by R

`plot(l1fit)`

↑ object returned by `glm`

R

" 2 "

Confidence Intervals

confint

Estimates with confidence intervals

	Estimate	Std. Error	Pr(> z )	2.5 %	97.5 %
(Intercept)	-75.79226904	8.658220532	2.062996e-18	-92.76206945	-58.82246862
year	<u>0.03117395</u>	0.004361182	8.801855e-13	<u>0.02262619</u>	<u>0.03972171</u>

log scale

Back-transformed estimates with confidence intervals

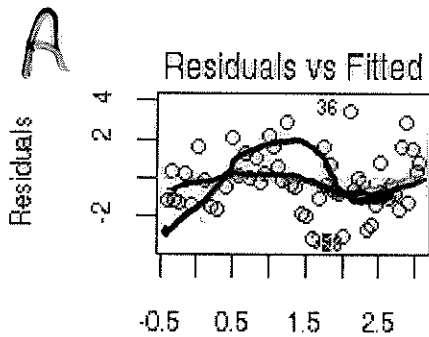
	Estimate	Std. Error	Pr(> z )	2.5 %	97.5 %
(Intercept)	1.212930e-33	8.658220532	2.062996e-18	5.175414e-41	2.842670e-26
year	<u>1.031665e+00</u>	0.004361182	8.801855e-13	<u>1.022884e+00</u>	<u>1.040521e+00</u>

exp. fit

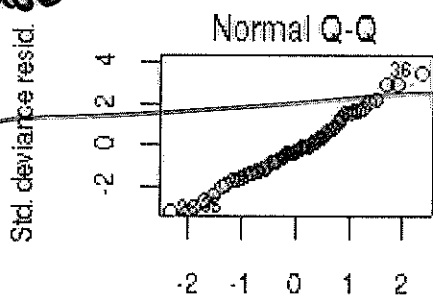
Back-transformed estimates with profile Confidence Intervals

	Estimate	Std. Error	Pr(> z )	2.5 %	97.5 %
(Intercept)	1.212930e-33	8.658220532	2.062996e-18	3.775357e-41	2.113763e-26
year	<u>1.031665e+00</u>	0.004361182	8.801855e-13	<u>1.023034e+00</u>	<u>1.040684e+00</u>

Examining Assumptions - see web page for function



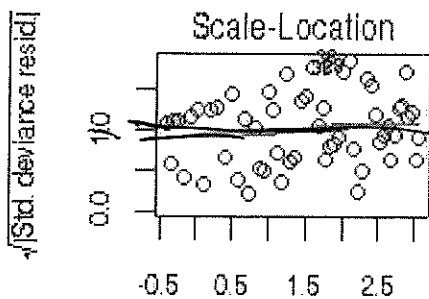
smoother (in red)



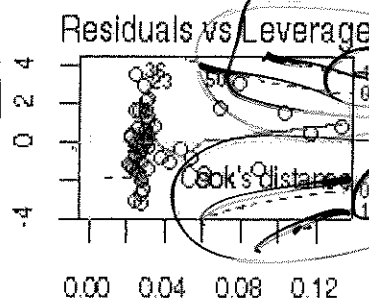
hint of non-normal

Predicted values  $\rightarrow \hat{\mu} = \log \mu$   
- log scale

par(mfrow = c(2,2))



Std. deviance resid.



2x2 arrange menu

Leverage

Plot  $\rightarrow$  generic function  
- based on residuals for  
linear model, adopted  
so works for generalized  
model

Plot (B)  $\rightarrow$  Normal Q-Q of  
standardized residuals  
- looks normal if  $\mu$ 's are  
big  
- not so helpful

Plot C - Plot of  $\sqrt{\text{standardized deviance residuals}}$   
 $\rightarrow$  predicted values

Recall - in regression

$$\rightarrow r_i = y_i - \hat{y}_i$$

$$\text{Var}(r_i) = \frac{\sigma^2}{n} (1 - h_{ii})$$

$$\sigma^2 (1 - h_{ii})$$

$h_{ii}$  is diagonal entry of  $\underbrace{X(X^T X)^{-1} X^T}_{H \text{ hat matrix}}$

$$\hat{y} = Hx \quad \hat{\varepsilon} = (I - H)y$$

$h_{ii} \rightarrow$  called leverage value  
 $\rightarrow$  potential for  $y_i$  to perturb  $\hat{\beta}$ , in case deletion diagnostic (if  $y_i$  removed)

eg.  $h_{ii} = 0$  dropping  $y_i$  leaves  $\hat{\beta}$  unchanged

to standardize residuals so have equal variances  
 define standardized residual

$$\tilde{r}_i = \frac{r_i}{\sqrt{1 - h_{ii}}} \quad \rightarrow \text{approximately } = \text{variances}$$

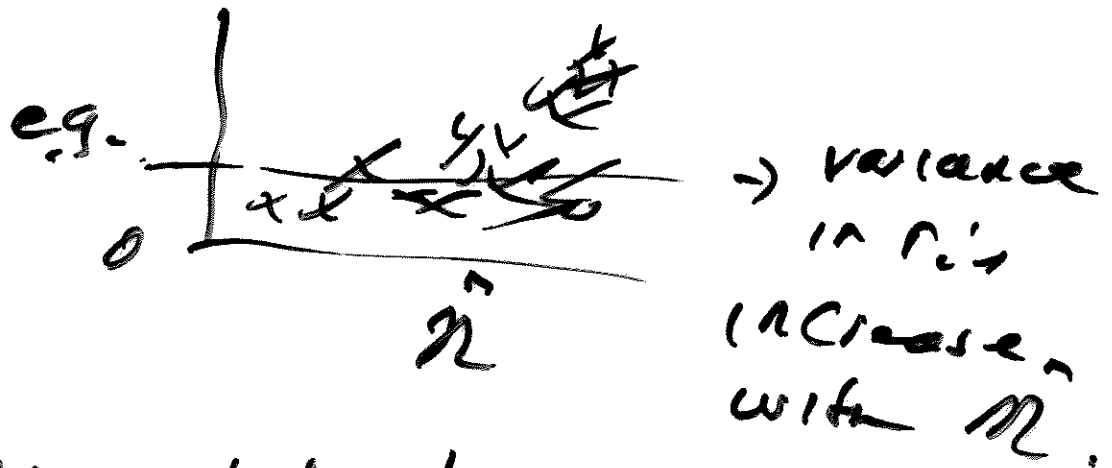
- do this for Pearson or

Deviance residual

Plot C  $\rightarrow$   $\sqrt{|\tilde{r}_i|}$  vs  $\hat{y}_i$

$\rightarrow$  used to examine patterns in dispersion

or variance



under model, dispersion should constant

Lastly,  $\hat{r}_i^2$  vs  $\hat{h}_{ii}$  - hat values,

$\otimes$  potential for infl. of  $y_i$  on fit

- if both big

$\rightarrow$   $i$ th observation is influential

- Cook's distance measures

the change in fitted values

$\hat{y}_i$  if  $y_i$  is deleted (dropped)

$\rightarrow$  obtained by examining  $\sum (\hat{y}_i - \hat{y}_i^{(i)})^2$

$\hat{y}_i^{(i)} \rightarrow$  fitted value for  $i$ th obsn if  $i$ th obsn is dropped

depends on  $\frac{(\hat{r}_i)^2 \times h_{ii}}{(p+1)(1-h_{ii})^2}$  = Cook's Distance

Influence  
Measure

- influential points in right hand corner of plot of  $\hat{r}_i$  vs  $h_{ii}$
- usual benchmarks are values .5 & 1

If Cook's distance = 1 or more point is fairly influential

ie Dropping  $y_i$  moves  $\hat{\beta}$  to edge of a 50% confidence region

In shark data - no overtly influential points

# Empirical model building

- Tentative model
- examine fit based on residual & other diagnostic values
- alter model to correct deficiencies

eg. potentially non-linear behaviour

$k \log \mu$  not linear in  $\beta_0 + \beta_1 x \dots \beta_k x^k$

- Remedies??

→ recall linear model case

- add more terms - non-linear

eg  $x^2$  etc.

- polynomial regression

- transform either  $y$  or  $x$ .

eg. by logs

log of  $y$  → also change variance pattern

in generalized linear models,  
tend not to transform  $y$ .

- interfere with link

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 - \beta_2 x_2$$

→ rather try different  $g$ 's

- practically - focus on

fixing  $x$ 's by transformation  
or most conveniently

by augmenting model new  
terms

eg. 3rd degree polynomial



Use of polynomials

- coefficients are not  
particularly interpretable

- need to look at  
pattern in graph