Generalized Linear Models (GLMs) with R

An Online Course

Presented by Geoffrey S. Hubona

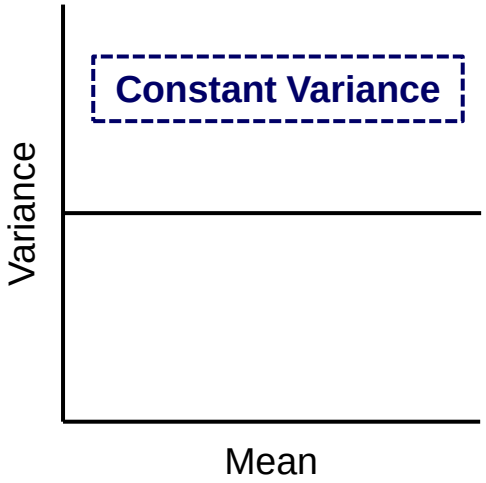
What Are GLMs?

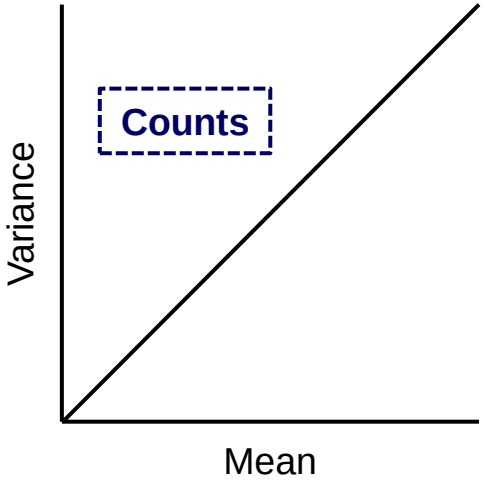
- **Linear models** (e.g. anova, manova, mancova, regression) have response variables (or more specifically, error terms) that are ‘well behaved’:
  - Constant variance at different mean levels of y.
  - Normally distributed error terms.
- Certain kinds of response variables invariably fail to achieve these lofty goals:
  - Count data expressed as proportions (e.g. logistic regressions).
  - Count data that are not proportions (e.g. log-linear count models).
  - Binary response variables (e.g. dead or alive).
  - Data on time to some event (e.g. time data with gamma errors).

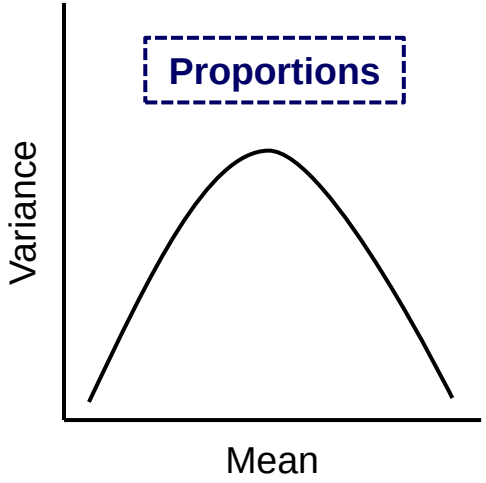
Three Properties of GLMs: (1) Error Structure

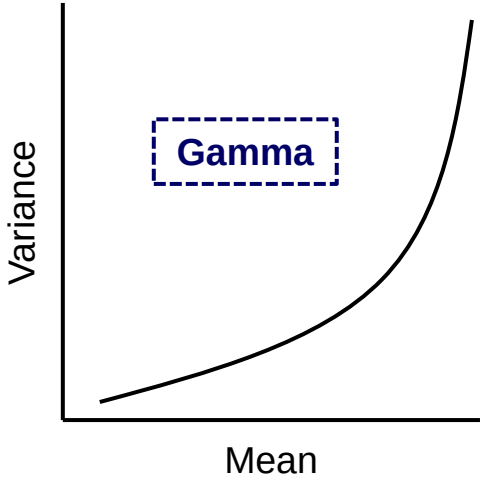
- ***Non-normal*** may mean errors that are: ***skewed***; ***kurtotic***; ***strictly bounded*** (as in proportions); cannot lead to ***negative fitted values*** (as in counts).
- GLMs allow the specification of different error distributions:
  - **Poisson** errors, useful with count data;
  - **Binomial** errors, useful with data on proportions;
  - **Gamma** errors, useful when there is a constant coefficient of variation; and
  - **Exponential** errors, typical in time-to-death (survival analysis).
- In R the **error structure** is defined by means of the **`family`** directive in the model formula, e.g. **`family = poisson`** or **`family = binomial`**.

Error Structures









Three Properties of GLMs: (2) Linear Predictor

- The structure of the model relates each observed y value to a ***predicted value***.
  - Predicted value is obtained by transforming the value emerging from the **linear predictor**.
- The linear predictor, η (eta), is the linear sum of the effects of one or more explanatory variables, x_j:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

- Where the xs are the values of the ρ explanatory variables, and the βs are the (usually) unknown parameters to be estimated.
- Right-hand side of equation is called the **linear structure**.

Three Properties of GLMs: (3) Link Function

- The **link function** relates the mean value of y to its linear predictor: $\eta = g(\mu)$
- The linear predictor, η, emerges from the linear model as the sum of the terms for each of the ρ parameters.
- This is not the value of y (except with the ***identity link***), it is the ***transformed value*** of y by the link function such that the predicted value of y is obtained by applying the ***inverse link function*** to η. Canonical link functions:

Error	Canonical Link
normal	*identity*
poisson	*log*
Binomial	*logit*
Gamma	*reciprocal*

More on Link Functions

- The most appropriate link function to use is the one that produces the ***minimum residual deviance***.
- Another important criterion for a link function is that the ***fitted values have reasonable bounds***:
  - Counts > 0 (use a *log link*)
  - 0 < proportions < 1 (use a *logit link*).
- Both **proportion data** (with binomial errors) and **count data** (with poisson errors) have at least three important properties:
  - Possible data values are bounded (see above);
  - Variance is non-constant (*humped* or *increasing* with mean);
  - Errors are non-normal.

Count Data

- There are at least two ways to estimate ***count data*** as the response variable using linear models:
- Count data as ***proportions***, where we know the number doing some particular thing, but we also know the number ***not*** doing that thing.
  - We assume that proportions have ***binomial errors*** and we use a ***logistic function*** (the ***logit link***) to model the 'connection' between the response variable and the independent variables.
- Count data as ***frequencies***, where we count how many times something happened, but we have no way of knowing how often it did ***not*** happen.
  - We assume that frequencies have ***poisson errors*** and use a ***logarithmic function*** (the ***log link***) to model the 'connection'.

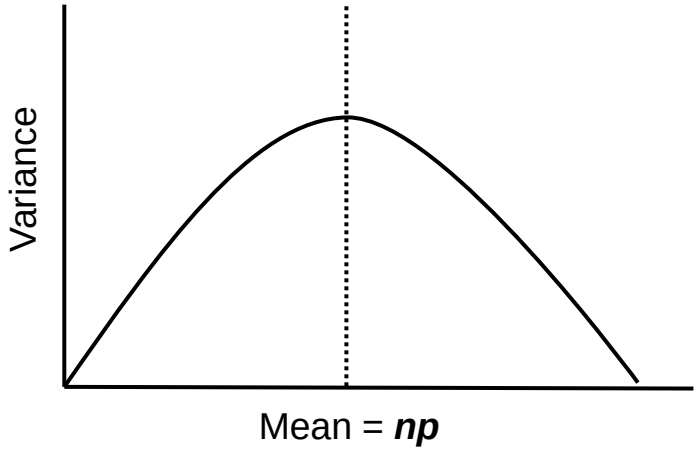
Proportion Data

- **Proportion data** (both p and q) is strictly bounded between ***zero*** and ***one***.
  - So if we used regression or ANCOVA, the fitted model could predict ***negative values*** or ***values > 1***.
- The **logistic** curve is often used to describe proportion data with p representing the proportion of individuals who respond in a certain way ("successes") and $1-p$ (or proportion q) representing the proportion who respond in other ways ("failures").
- A third variable that is relevant is the size of the sample, n, from which p was estimated which represents the ***number of attempts***.

Binomial Errors

- In probability theory and statistics, the ***binomial distribution*** is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields a probability of success p.
- The ***variance*** for the binomial distribution is not constant with mean np is: $s^2 = np(1-p)$ or $s^2 = npq$ where q is the probability of failure.

Proportions: Binomial Distribution



Proportions and Odds

- What is the distinction between the ***probability*** of some event occurring and the ***odds*** of the event?
- A reasonably good horse who wins 2/3 of the races s/he enters: **p(winning) = 0.6667**. The probability is the ratio:
  - ***# successful trials / # total trials***.
- However, a bookmaker would tell you that the ***odds*** of the horse winning are **2 to 1**. The odds are:
  - ***# successful trials / # unsuccessful trials.***

How does this Relate To Binomial Distributions?

- We said the **odds** is the probability of success p divided by the probability of failure q:

$$\frac{p}{q} = \frac{e^{a+bx}}{1+e^{a+bx}}\left[1-\frac{e^{a+bx}}{1+e^{a+bx}}\right]^{-1} \boxed{= e^{a+bx}}$$

- Taking natural logs and recalling that $\ln(e^x) = x$ leaves us with:

$$\ln(p/q) \boxed{= a+bx}$$

- This yields a **linear predictor**, $a+bx$ not for p but for the ***logit*** transformation of p, namely $\ln(p/q)$.
  - In **R**, the logit is the link function relating the linear predictor to the value of p.

Why Not Perform a Linear Regression?

- Instead of doing implicit transformations from y ot the x variables through a link function, why not simply do a linear regression of $\ln(p/q)$ against the explanatory x variable?

1) **R** allows for the non-constant binomial variance
2) **R** knows and deals with the fact that logits for p's near 0 or 1 are infinite.
3) **R** uses weighted regression to allow for differences between the sample sizes.

Binomial Models and Heart Disease

- Early diagnosis of heart disease is critical.
- One diagnostic aid is the level of enzyme ***creatinine kinase*** (CK) in the blood.
- Study (Smith, 1967) looked at level of CK for 360 patients thought to have had a heart attack.
- Data is on the next slide.
- ***Can we estimate the probability that a patient has had a heart attack using CK level?***

Binomial Models and Heart Disease

CK Value	Patients with Heart Attack	Patients without Heart Attack
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

data.frame **`heart`** with variables **`ha`**, **`ok`** and **`ck`**

Modeling the GLM in R Script

```
> heart <- read.csv("c:/temp/heart.csv",header=T)
> heart # to view the entire data set
    ck ha ok
1   20  2 88
2   60 13 26
3  100 30  8
4  140 30  5
5  180 21  0
6  220 19  1
7  260 18  1
8  300 13  1
9  340 19  1
10 380 15  0
11 420  7  0
12 460  8  0
```

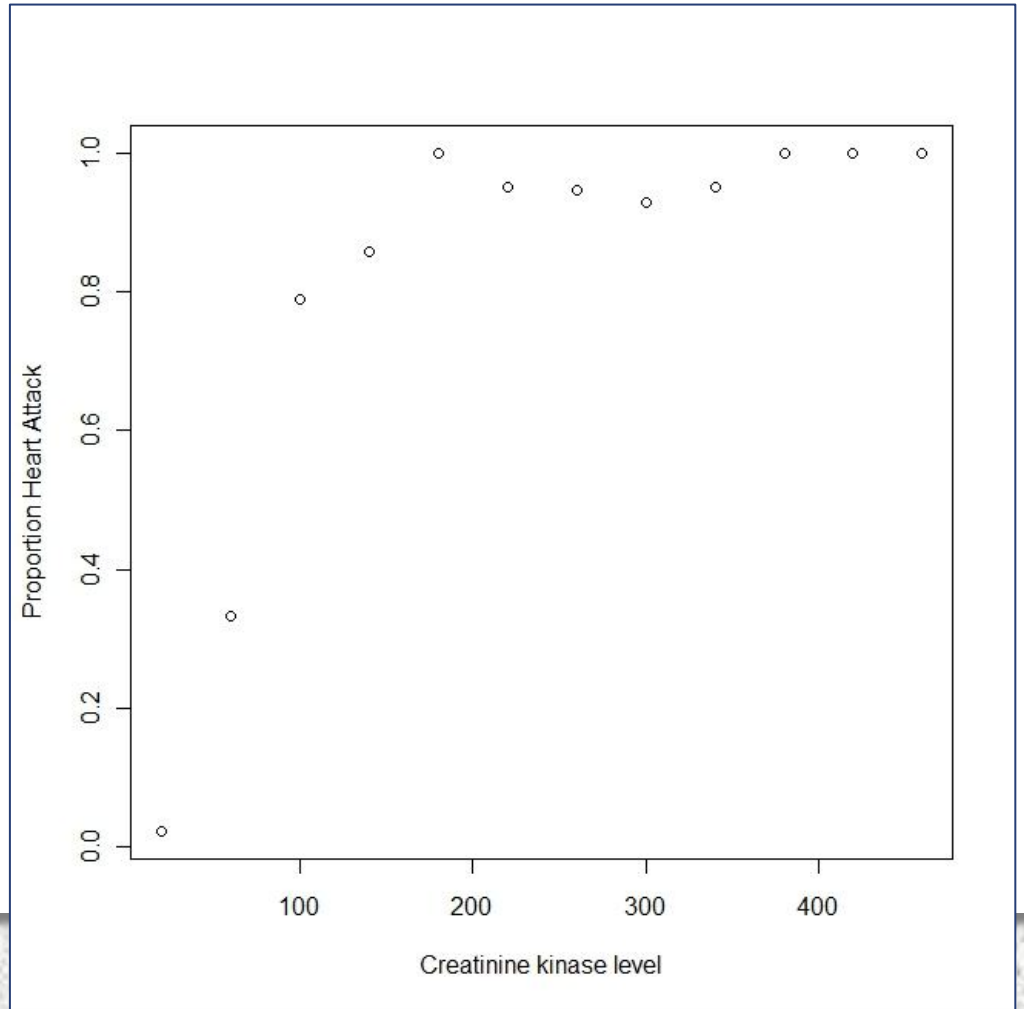
Calculate Proportions By CK Level

```
> p <- heart$ha /(heart$ha+heart$ok)
> p
 [1] 0.02222 0.33333 0.78947 0.85714 1.00000 0.95000
 [7] 0.94737 0.92857 0.95000 1.00000 1.00000 1.00000
```

Proportions of patients who suffered a heart attack at each CK level

Plot Proportions By CK Level

```
> plot(heart$ck,p,xlab="Creatinine kinase level",
  ylab="Proportion Heart Attack")
```



Proportions By CK Level

- Expected value of proportions can be specified as:

$$E(p_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \qquad (1)$$

- So expected number of heart attack sufferers is:

$$\mu_i \equiv E(p_i N_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} N_i \qquad (2)$$

- Where N_i is total number of patients at each CK level.

Proportions By CK Level

- So the model is non-linear in its parameters:

$$g(\mu_i) = \log\left(\frac{\mu_i}{Ni - \mu_i}\right) \qquad (3)$$

- But when we apply the ‘logit’ link:

$$g(\mu_i) = \beta_0 + \beta_1 x_i \qquad (4)$$

- The right hand side is linear in its model parameters.

Binomial Model Specification in R

Two ways to specify a binomial model in R:

- The response variable can be ***observed proportion of successful binomial trials***:
  - Must supply number of trials in weights argument to **`glm`**. For binary data, no weights vector is needed as 1 is the default weights.
- Response variable can be ***supplied as a two column array***, such that the first column indicates the number of binomial '***successes***' and second column provides the number of binomial '***failures***'.

Let's illustrate the second one now !

Binomial Model Specification in R

- We supply two arrays on the r.h.s. of the model formula using the R function **`cbind()`**
- Here is a **`glm`** call which will fit the heart attack model:

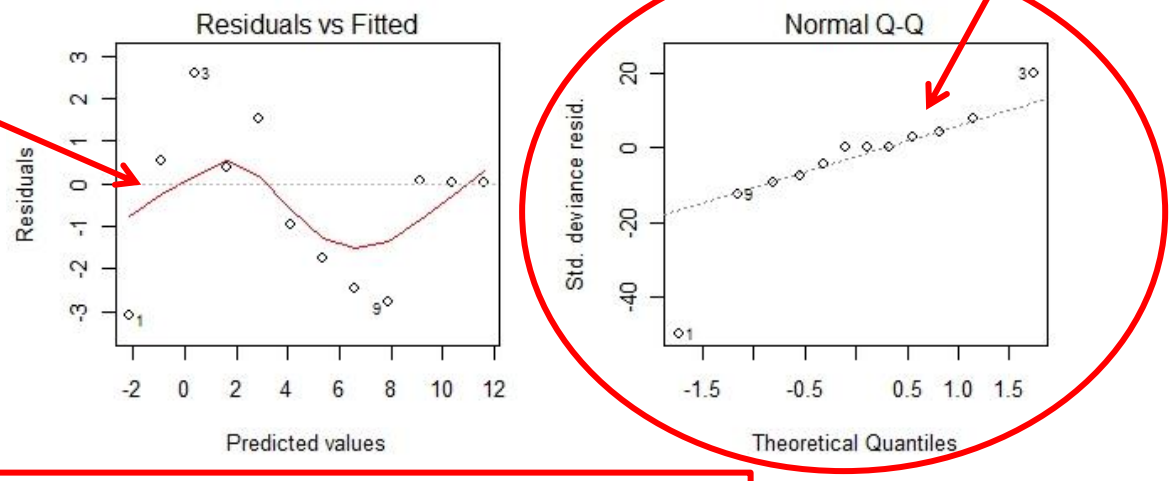
```
> mod.0 <- glm(cbind(ha,ok)~ck, family=binomial
  (link=logit),data=heart)
> mod.0

Call: glm(formula = cbind(ha, ok) ~ ck,
family = binomial(link = logit),data = heart)
Coefficients:
(Intercept)            ck
   -2.75836       0.03124
Degrees of Freedom: 11 Total (i.e. Null);  10 Residual
Null Deviance:      271.7
Residual Deviance: 36.93      AIC: 62.33
```

Residual Plots

Some departure from straight line is expected

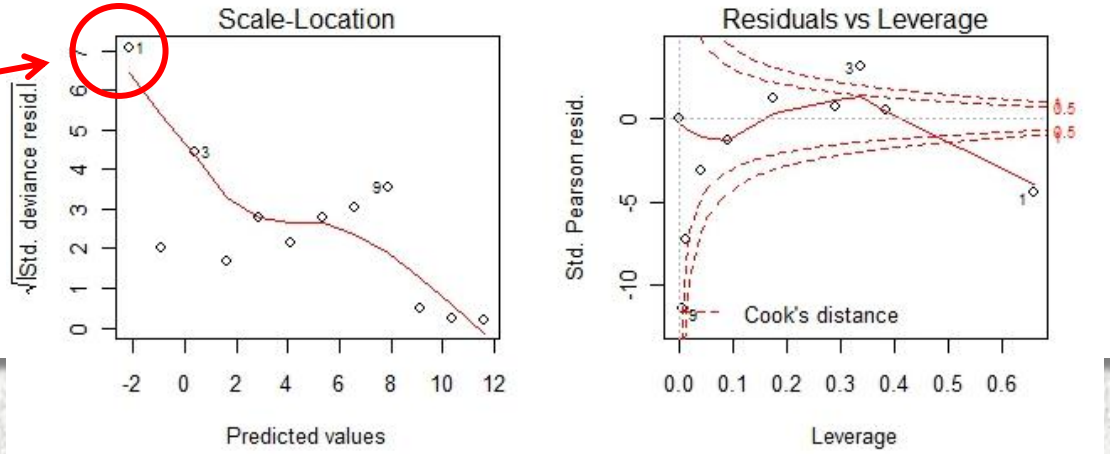
```
> par(mfrow=c(2,2))
> plot(mod.0)
```

Suggests a cubic function



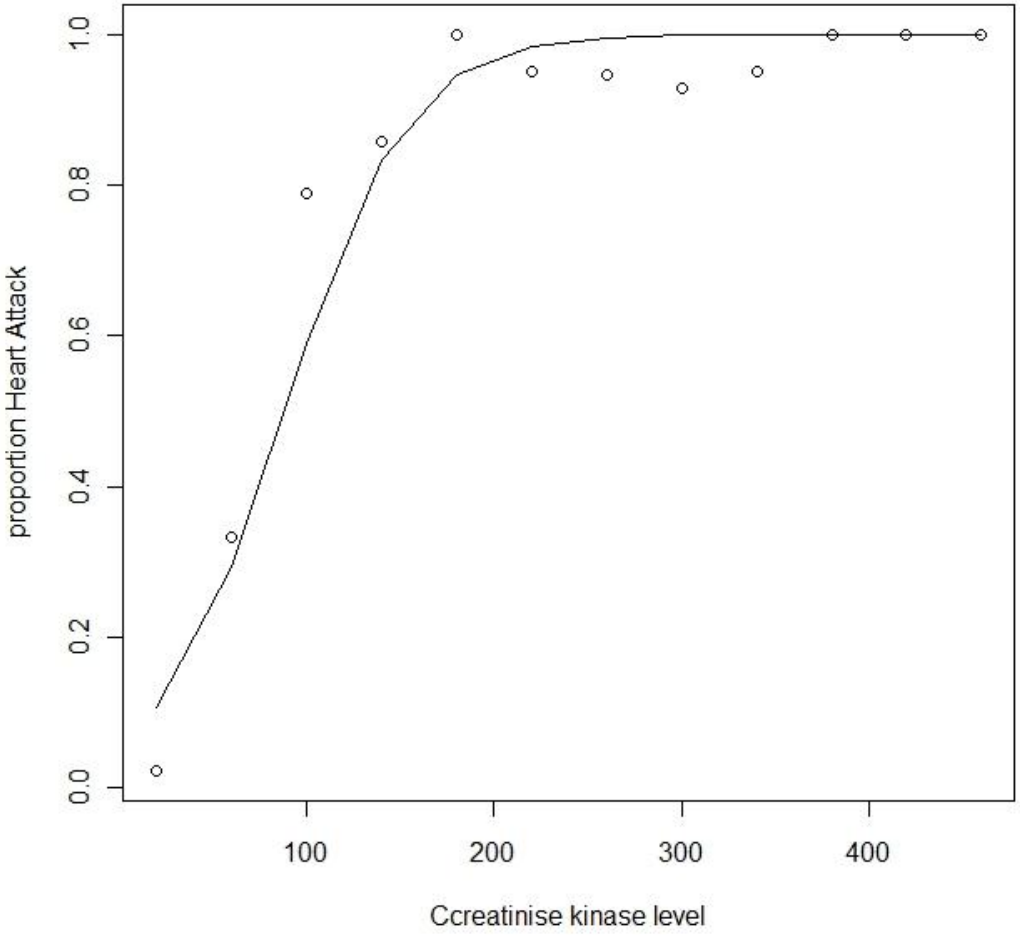
Predicted values are on the scale of linear predictor

High influence



Plot of Predicted Heart Attack by CK Level

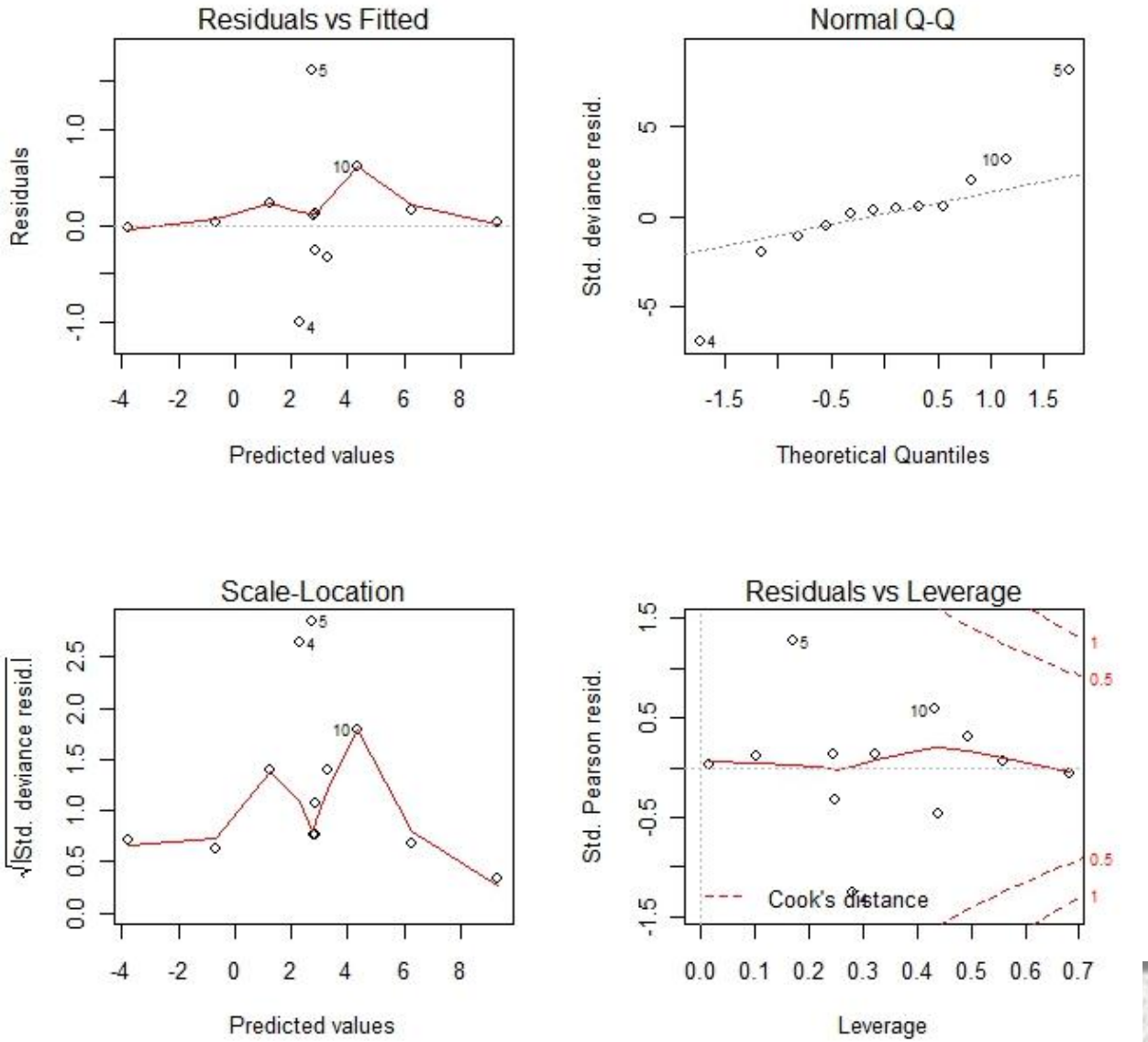
```
> plot(heart$ck,p,xlab="Creatinine Kinase level"
  ylab="Proportion Heart Attack")
> lines(heart$ck,fitted(mod.0))
```



Cubic Linear Predictor

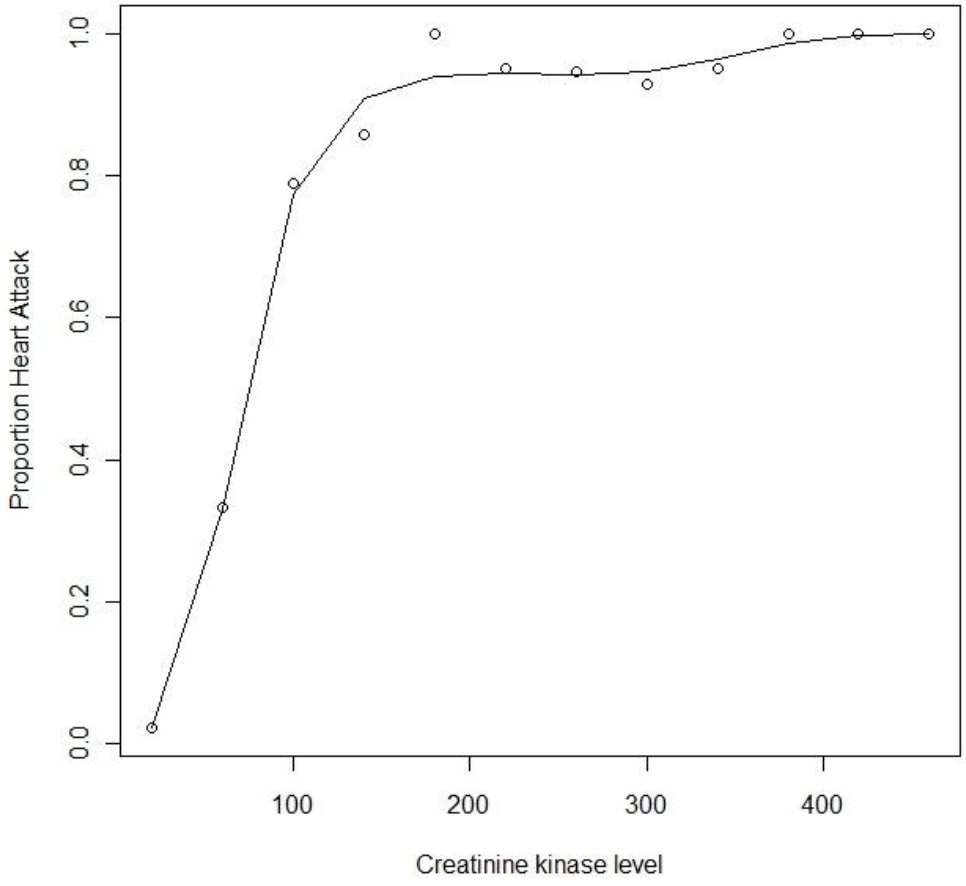
```
> mod.2 <- glm(cbind(ha,ok)~ck+I(ck^2)+I(ck^3),
  family=binomial, data=heart)
> mod.2

Coefficients:
(Intercept)            ck       I(ck^2)       I(ck^3)
 -5.786e+00    1.102e-01   -4.649e-04    6.448e-07

Degrees of Freedom: 11 Total (i.e. Null);  8 Residual
Null Deviance:      271.7
Residual Deviance: 4.252      AIC: 33.66
```

Residual Plots for Second Cubic Model



Second Plot of Predicted Heart Attack by CK Level

```
> par(mfrow=c(1,1))
> plot(heart$ck,p,xlab="Creatinine Kinase level"
  ylab="Proportion Heart Attack")
> lines(heart$ck,fitted(mod.2))
```

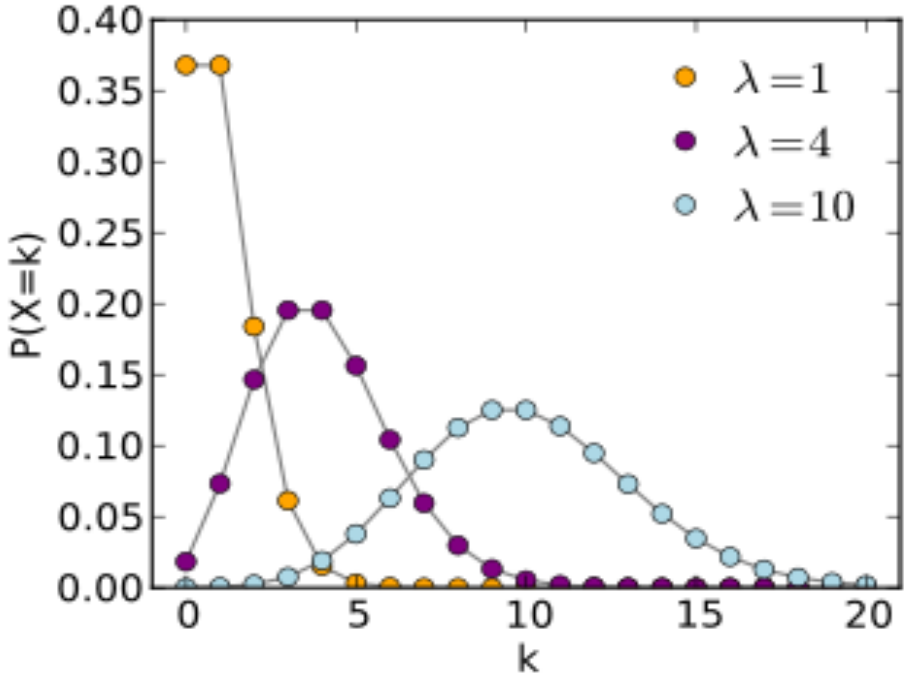


Second Plot of Predicted Heart Attack by CK Level

```
> par(mfrow=c(1,1))
> plot(heart$ck,p,xlab="Creatinine Kinase level"
  ylab="Proportion Heart Attack")
> lines(heart$ck,fitted(mod.2))

> anova(mod.o,mod.2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ha, ok) ~ ck
Model 2: cbind(ha, ok) ~ ck + I(ck^2) + I(ck^3)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        10     36.929
2         8      4.252  2   32.676 8.025e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
```

Count Data as Frequencies: Poisson

- In probability theory and statistics, the ***Poisson distribution*** is a discrete probability distribution that expresses the probability of a given number of independent events occurring in a fixed interval of time and/or space and with a known average rate of occurrence.



Count Data as Frequencies: Poisson

- If λ is the expected number of occurrences in a given interval k, then the probability that there are exactly k occurrences is equal to:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

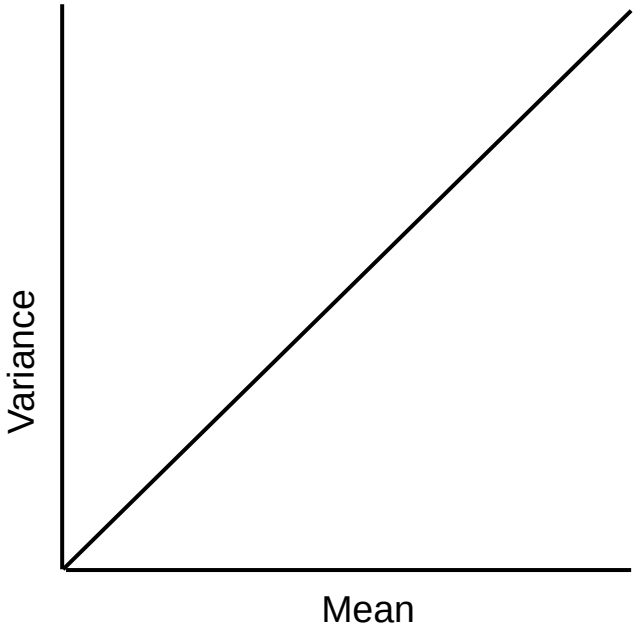
where:

e is the base of the natural logarithm ($e = 2.7182\ldots$)
k is the number of occurrences of an event
$k!$ is the factorial of
λ is a positive real number

Poisson Errors

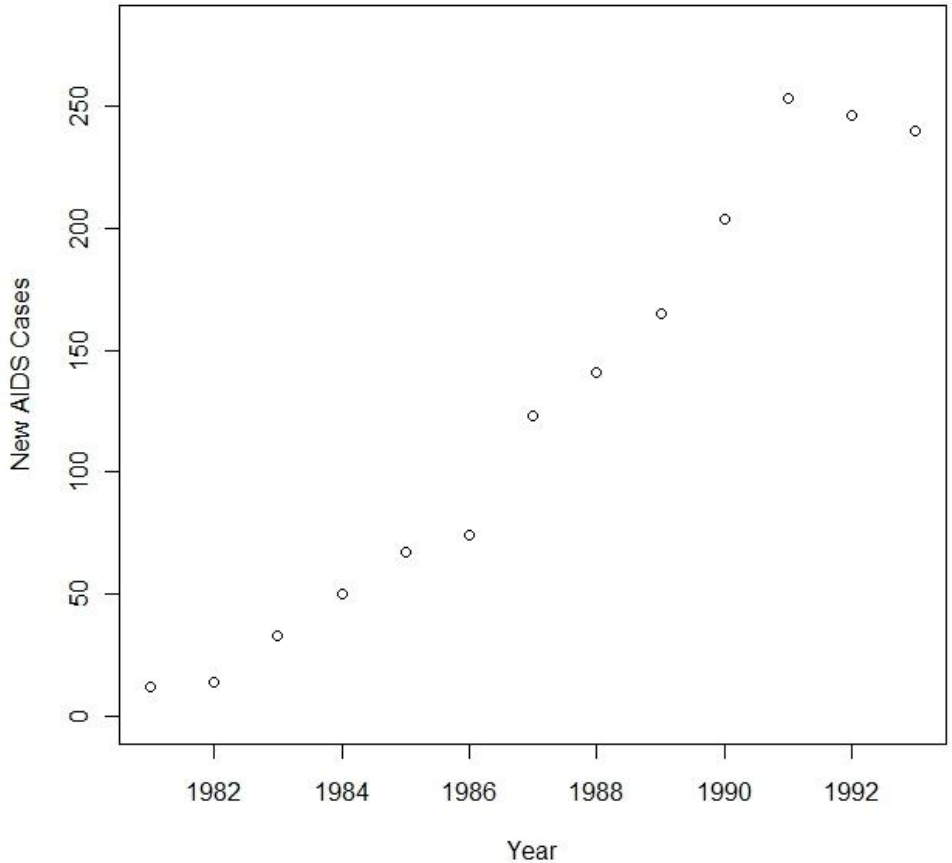
- Linear regression methods (assume constant variance and normal errors) are not appropriate for count data:
  - Possible to predict negative counts with linear regression
  - Variance of response variable increases with the mean
  - Errors will not be normally distributed
  - Zeros are a headache in transformations

Counts as Frequencies: Poisson Distribution



A Poisson Regression Epidemic Model

```
> y <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
> t <- 1:13
> plot(t+1980,y,xlab="Year",ylab="New AIDS
  Cases",ylim=c(0,280))
```



Data provided by Venables and Ripley, 2002

AIDS Epidemic Poisson Regression Model

- Model assumes that number of new cases per year:

$$\mu_i = \gamma \exp(\delta t_i) \qquad (5)$$

- Where δ and γ are unknown, and t_i is time in years since the start of the data.
- A log link turns this into a GLM:

$$\log(\mu_i) = \log(\gamma) + \delta t_i = \beta_0 + t_i \beta_1 \qquad (6)$$

- And we assume that $y_i \sim Poi(\mu_i)$ where y_i is the observed number of new cases.

Fit a Poisson Model

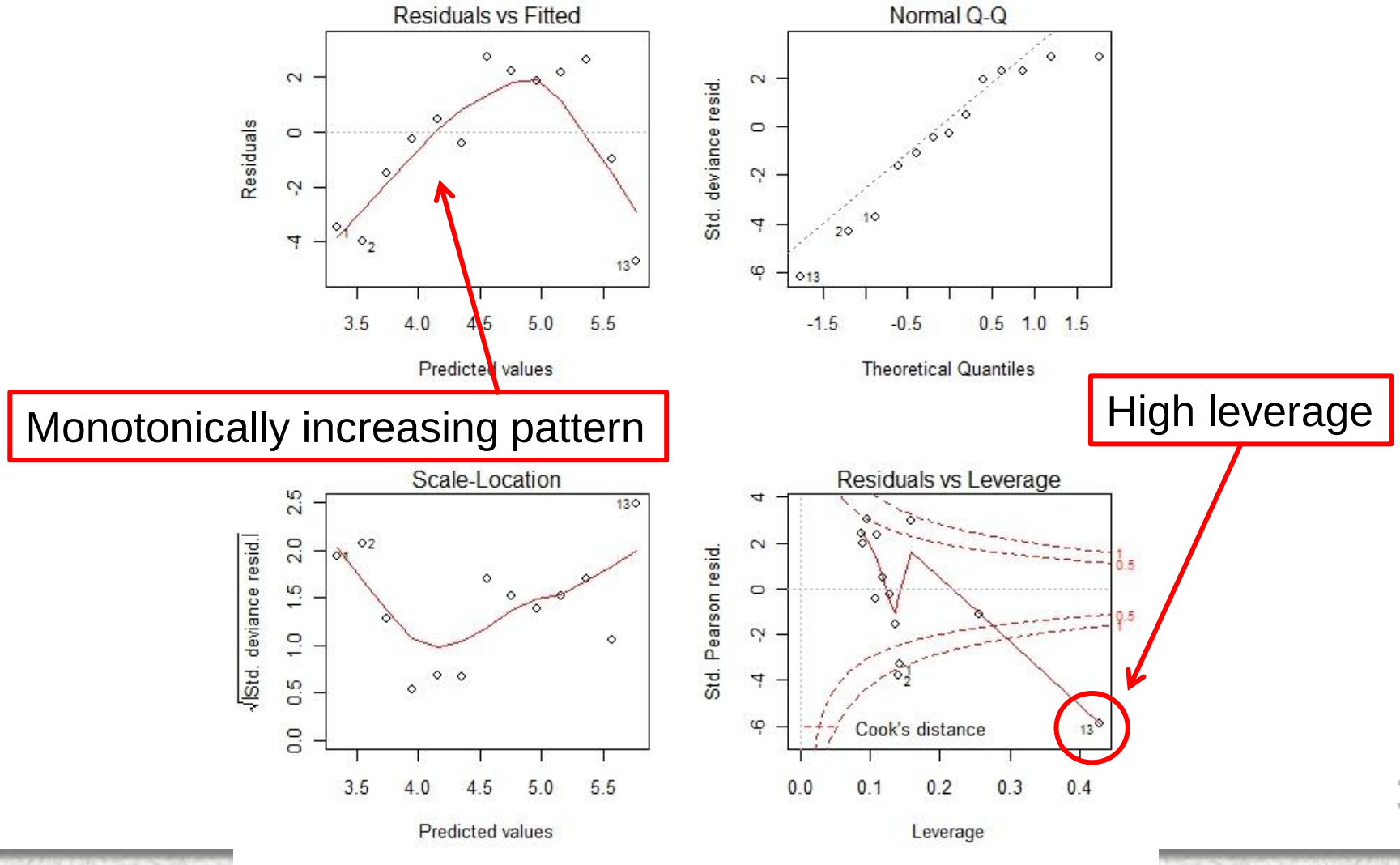
```
> m0 <- glm(y~t,poisson)
> m0

Call:  glm(formula = y ~ t, family = poisson)

Coefficients:
(Intercept)            t
     3.1406       0.2021

Degrees of Freedom: 12 Total (i.e. Null);  11
Residual
Null Deviance:          872.2
Residual Deviance: 80.69      AIC: 166.4
```

Deviance too high for random variable with 11 d.o.f.

AIC too high

Residual Plots for Poisson AIDS Model

```
> par(mfrow=c(2,2))
> plot (m0)
```



Add a Quadratic (Time) Term to Poisson AIDS Model

- We add a quadratic term to the model:

$$\mu_i = \exp\left(\beta_0 + \beta_1 t_i + \beta_2 t_i^2\right) \qquad (7)$$

- This model allows situations other than the unrestricted spread of the disease to be represented.
- We fit the model and check it on the following slide.

Fit the Quadratic Poisson Model

```
> m1 <- glm(y~t+I(t^2),poisson)
> plot(m1)
> summary(m1)
Call:
glm(formula = y ~ t + I(t^2), family = poisson)
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 ***
t            0.556003   0.045780  12.145  < 2e-16 ***
I(t^2)      -0.021346   0.002659  -8.029 9.82e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 872.2058  on 12  degrees of freedom
Residual deviance:   9.2402  on 10  degrees of freedom
AIC: 96.924
Number of Fisher Scoring iterations: 4
```

Time squared

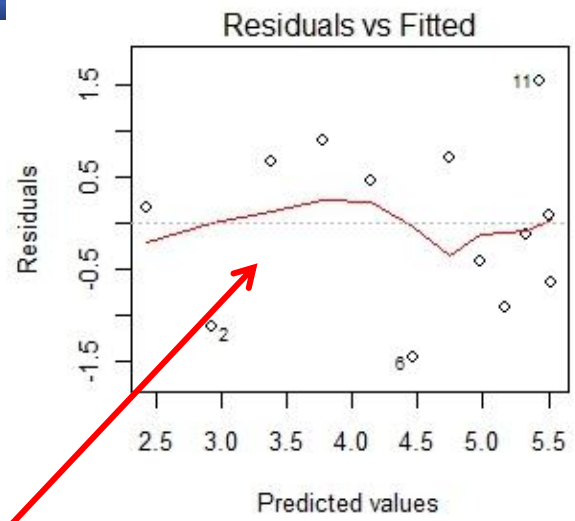
Reasonable deviance for random variable with 10 d.o.f.

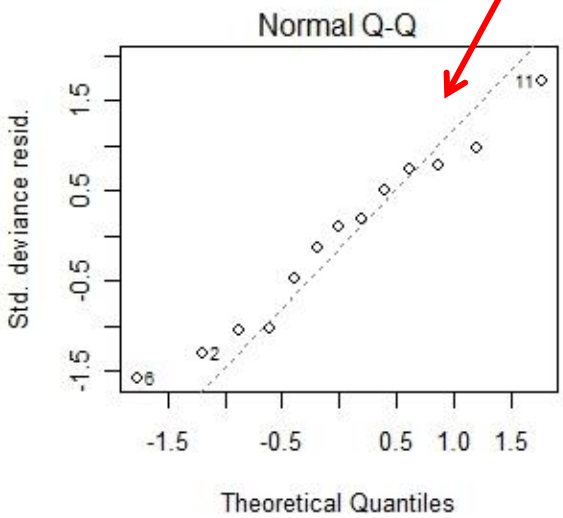
AIC much reduced

Residual Plots for Quadratic AIDS Model

QQ line is straighter

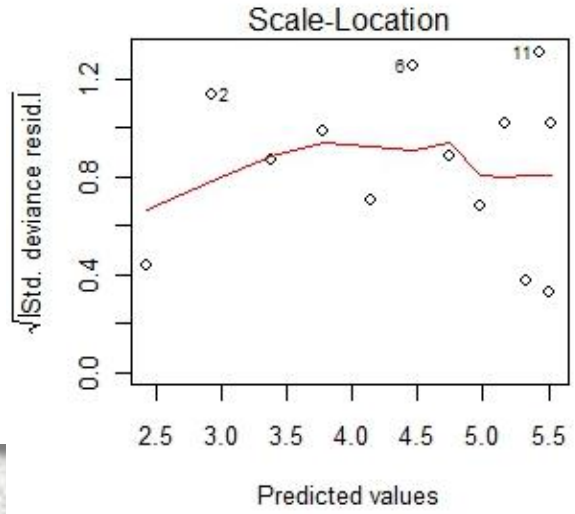
```
> plot(m1)
```

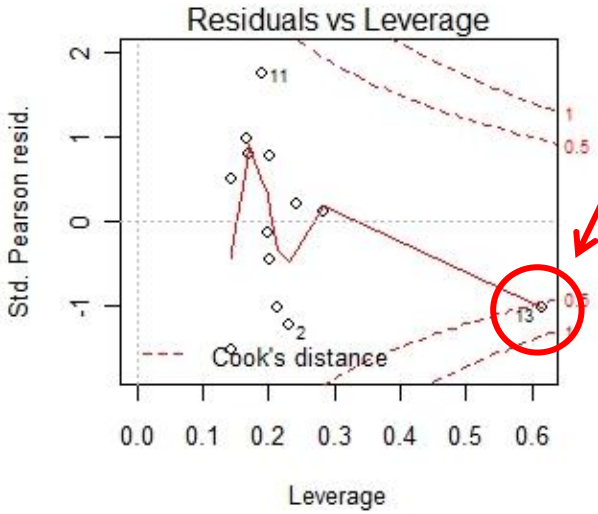




No noticeable pattern

Leverage Reduced





What Does Coefficient β_1 Represent ?

First Model

```
Call:  glm(formula = y ~ t, family = poisson)

Coefficients:
(Intercept)            t
     3.1406       0.2021
```

(Exponentiated) Uncon-trolled Spread of AIDS

Second Model (Quadratic termed added)

```
Call: glm(formula = y ~ t + I(t^2), family = poisson)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.901459   0.186877  10.175  < 2e-16 ***
t             0.556003   0.045780  12.145  < 2e-16 ***
I(t^2)       -0.021346   0.002659  -8.029 9.82e-16 ***
```

β_1 is greater in the more complex, but better, model

Point Estimate of Confidence Interval for β_1

```
> beta.1 <- summary(m1)$coefficients[2,]
> ci <- c(beta.1[1]-
  1.96*beta.1[2],beta.1[1]+1.96*beta.1[2])
> ci # print 95% CI for beta_1

> ci
 Estimate  Estimate
0.4662750 0.6457316
```

95% Confidence Interval for β_1 in Model #2

Estimate of Confidence Interval for β_1 Over Time

```
> new.t <- seq(1,13,length=100)
> fv <- predict(m1,data.frame(t=new.t),se=TRUE)
> plot(t+1980,y,xlab="Year",ylab="New AIDS
  Cases",ylim=c(0,280))
> lines(new.t+1980,exp(fv$fit))
> lines(new.t+1980,exp(fv$fit+2*fv$se.fit),lty=2)
> lines(new.t+1980,exp(fv$fit-2*fv$se.fit),lty=2)
```

Confidence Interval At Each Point in Time