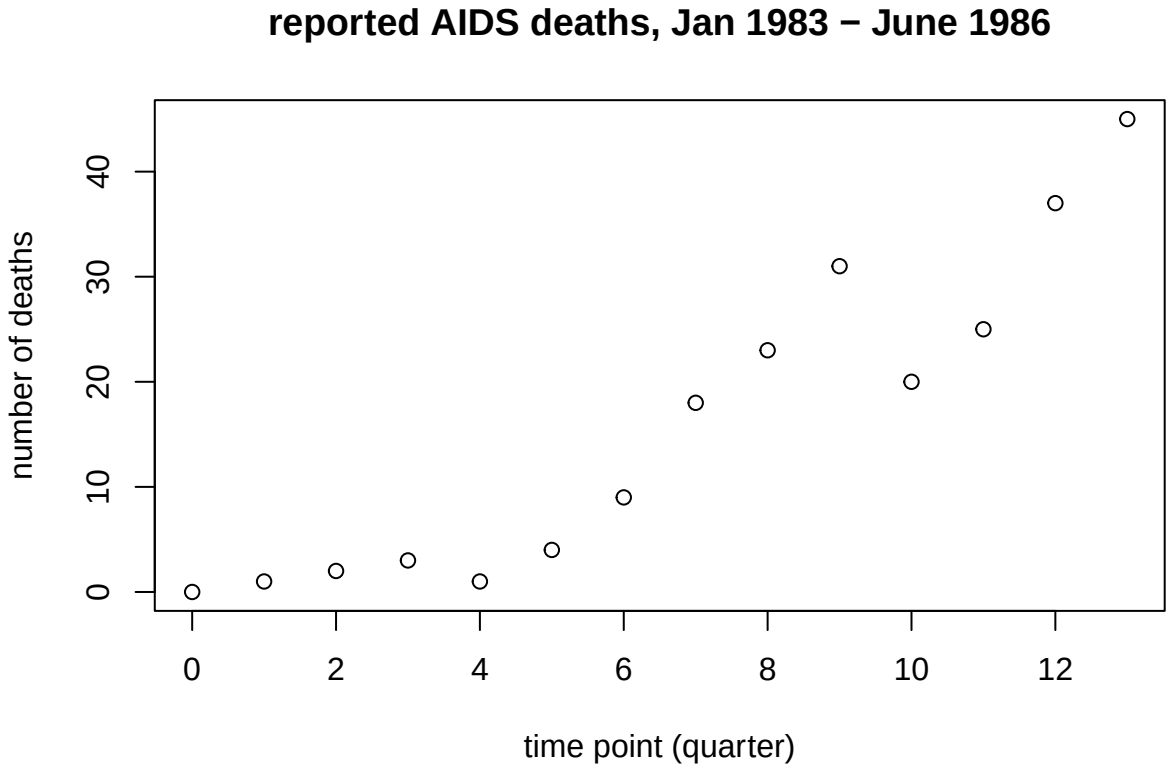
Introduction to Poisson regression

2025-05-24

Whyte *et al* (1987) developed a model for the number of AIDS cases in Australia, using quarterly AIDS incidence beginning January, 1983, through December 19, 1986. A subset of the data, comprising quarterly cases from January, 1983 through end of June, 1986, is included in the book on generalized linear models by Dobson and Barnett (2018). It is shown in the R code and plot below.

```
x=0:13
y=c(0,1,2,3,1,4,9,18,23,31,20,25,37,45)
plot(x,y,ylab="number of deaths",xlab="time point (quarter)",
 main="reported AIDS deaths, Jan 1983 - June 1986")
```



The outcome variable y takes non-negative integer values, and so a linear regression model of y on x makes no sense, as the regression model assumes an outcome variable which is normally distributed, so real valued and not constrained to be non-negative.

One choice of distribution for non-negative integer random variables is the Poisson distribution. If Y has a Poisson distribution with mean λ, then

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \ y = 0, 1, \ldots$$

where $\lambda > 0$. If the mean of Y is made to depend on x, this can be extended to

$$P(Y = y|X = x) = \frac{\lambda(x)^y e^{-\lambda(x)}}{y!},\ y = 0, 1, \ldots$$

$\lambda(x)$ must be non-negative. The Poisson regression model accomplishes this by setting

$$log(\lambda(x)) = \beta_0 + \beta_1 x$$

so that $\lambda(x) = e^{\beta_0+\beta_1 x}$, which is non-negative.

The right hand side of the equation for $log(\lambda(x))$ is the same as that of a linear regression. Here it's a simple linear regression, but with additional predictor variables the right hand side can be extended to $\beta'\mathbf{x}$.

Generalized linear models always have a common structure where some function of the object of interest (here $log(\lambda(x))$) is a linear function of the predictor variables.

This model can be fit in R using the "glm" procedure, specifying a Poisson distribution, as follows.

```
preg.out=glm(y~x,family=poisson)
summary(preg.out)
```

```
##
## Call:
## glm(formula = y ~ x, family = poisson)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.21008  -1.02032  -0.69704   0.04028   2.70758
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.59616    0.23004   2.592  0.00955 **
## x            0.25652    0.02204  11.639  < 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: 207.272  on 13  degrees of freedom
## Residual deviance:  29.654  on 12  degrees of freedom
## AIC: 86.581
##
## Number of Fisher Scoring iterations: 5
```

As with the logistic regression model, we won't try to understand details of the estimation and testing results, which are discussed at length in Stat4620. Suffice it to say that in this example, x is highly significant (p-value $< .001$).

The estimate of Y when the predictor takes value x is given by the estimated mean

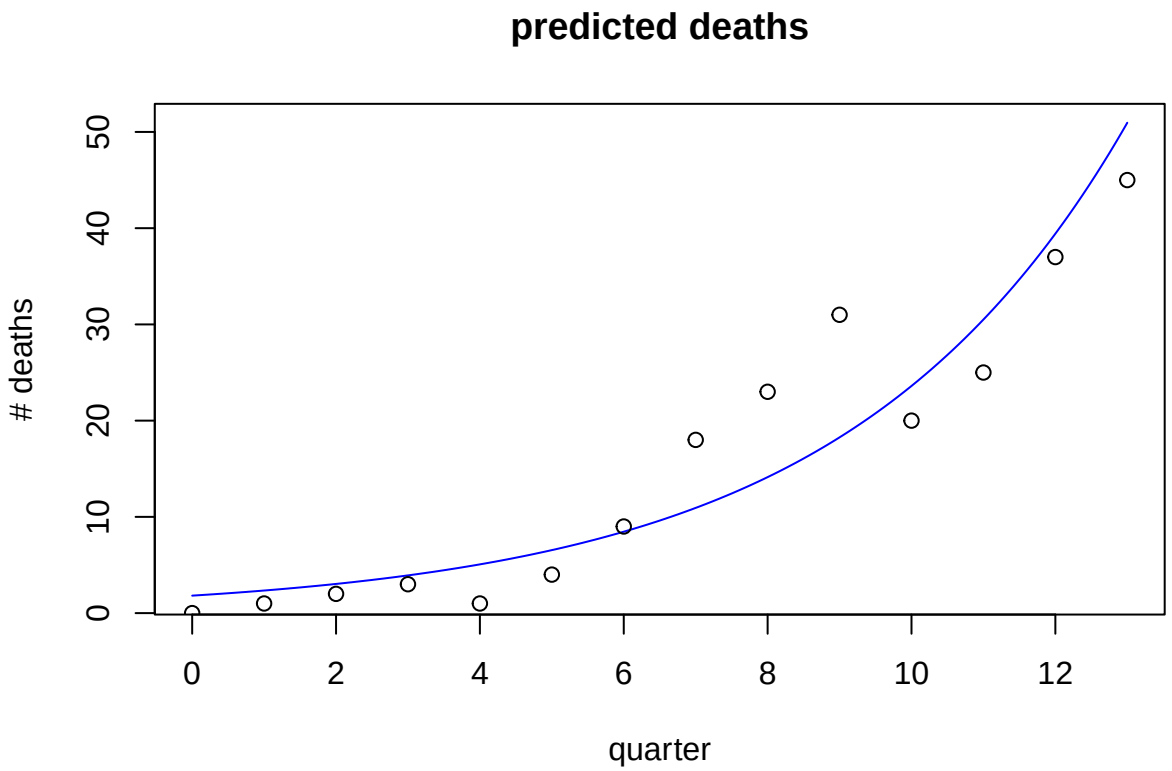
$$\hat{\lambda}(x) = e^{\hat{\beta}_0+\hat{\beta}_1 x}$$

where from the output above $\hat{\beta}_0 \approx .596$ and $\hat{\beta}_1 \approx .257$.

The estimate is shown as the line in the following plot, which has the data superimposed.

```
xtick=seq(0,13,length.out=500)
lambdahat=exp(coef(preg.out)[1]+coef(preg.out)[2]*xtick)
```

```
plot(xtick,lambdahat,col="blue",type="l",xlab="quarter",ylab="# deaths",main="predicted deaths")
points(x,y)
```



There appears to be a discontinuity in the data at 9-10 months which would merit further investigation.

References:

Bruce M. Whyte, Julian Gold, Annette J. Dobson, and David A. Cooper. Epidemiology of acquired immunodeficiency syndrome in Australia. Med J Aust 1987; 146: 63-9.

Dobson, A.J. and Barnett, A.G. (2018) An introduction to generalized linear models, Chapman-Hall.