Event Analysis

Event Counts and Event History Analysis

David Barron

Hilary Term 2017

Introduction

Outline

- Count data methods
  - Contingency tables
  - Poisson regression
  - Negative binomial regression
- Event History Analysis
  - Basic Concepts
  - Continuous time models
  - Discrete time models

Contingency Tables

2-dimensional contingency tables

The simplest way of dealing with counts is to create a *contingency table*. Here is a very simple example:

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|-------------------------|

============================================================
                  Number of children
Works part time          0        1        2       3   Total
------------------------------------------------------------
no                   29546     3405      868      53   33872
                    29141.4   3747.5    932.8    50.3
                      5.619    31.307    4.503   0.145
------------------------------------------------------------
yes                   9848     1661      393      15   11917
                    10252.6   1318.5    328.2    17.7
                     15.970    88.986   12.800   0.411
------------------------------------------------------------
Total                39394     5066     1261      68   45789
============================================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 160      d.f. = 3      p <2e-16
```

Tests for association

If there is no association between the two variables (i.e., they are independent), then the probability of a case falling into category row i and category column j is just determined by the probability of being in category i multiplied by the probability of being in category j.

$$\pi_{ij} = \pi_{i+}\pi_{+j}.$$

Testing independence

We can convert this to a count by multiplying by the total number of cases in the table, n. We can test the goodness of fit of this model, the null hypothesis being that any divergence from a perfect fit is due only to sampling error. Pearson's chi-squared statistic is:

$$\chi^2 = \sum\sum \frac{(m_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}},$$

where m_{ij} is the observed count in cell ij, and $\hat{m}_{ij}$ is the expected count under the null hypothesis.

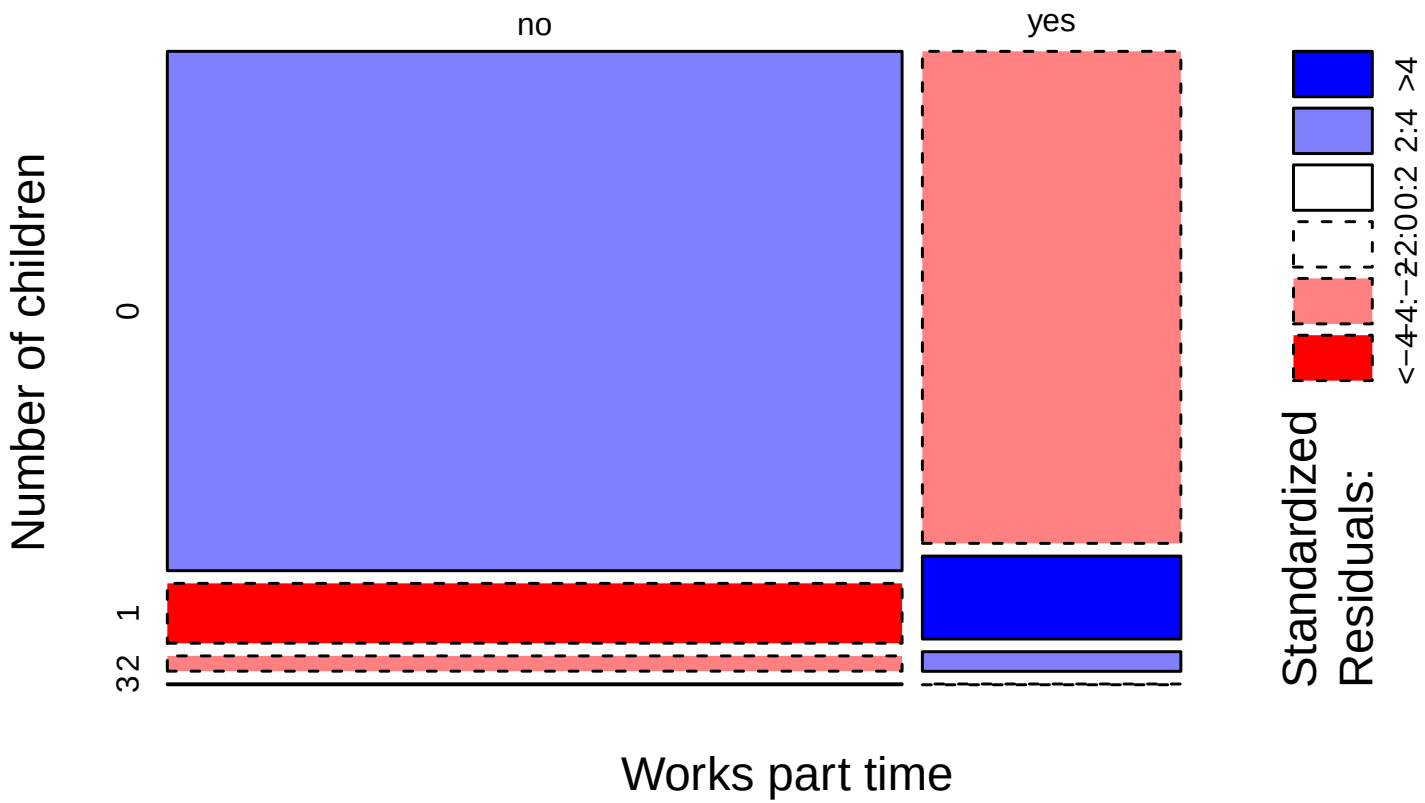
Alternatively, one can use a likelihood ratio chi-square (G^2):

$$G^2 = 2\sum\sum m_{ij} \log(m_{ij}/\hat{m}_{ij}).$$

In both cases there are $(I-1)(J-1)$ degrees of freedom, where I is the number of rows and J is the number of columns in the table, respectively.

Extensions

It is possible to extend this basic idea to higher dimensions of tables. There are also methods that can be used for ordinal variables. Graphical methods are often useful for showing more detail about the nature of the association.



Count regression methods

Poisson regression

Counts of events

It is quite common to be faced with a requirement to analyse data that are *counts* of the occurrence of some event. Typical examples are number of new entries in a market, number of new rules created in an organization, number of job titles, visits to the doctor, number of complaints received, and many more. This implies that the variable can take only non-negative integer values.

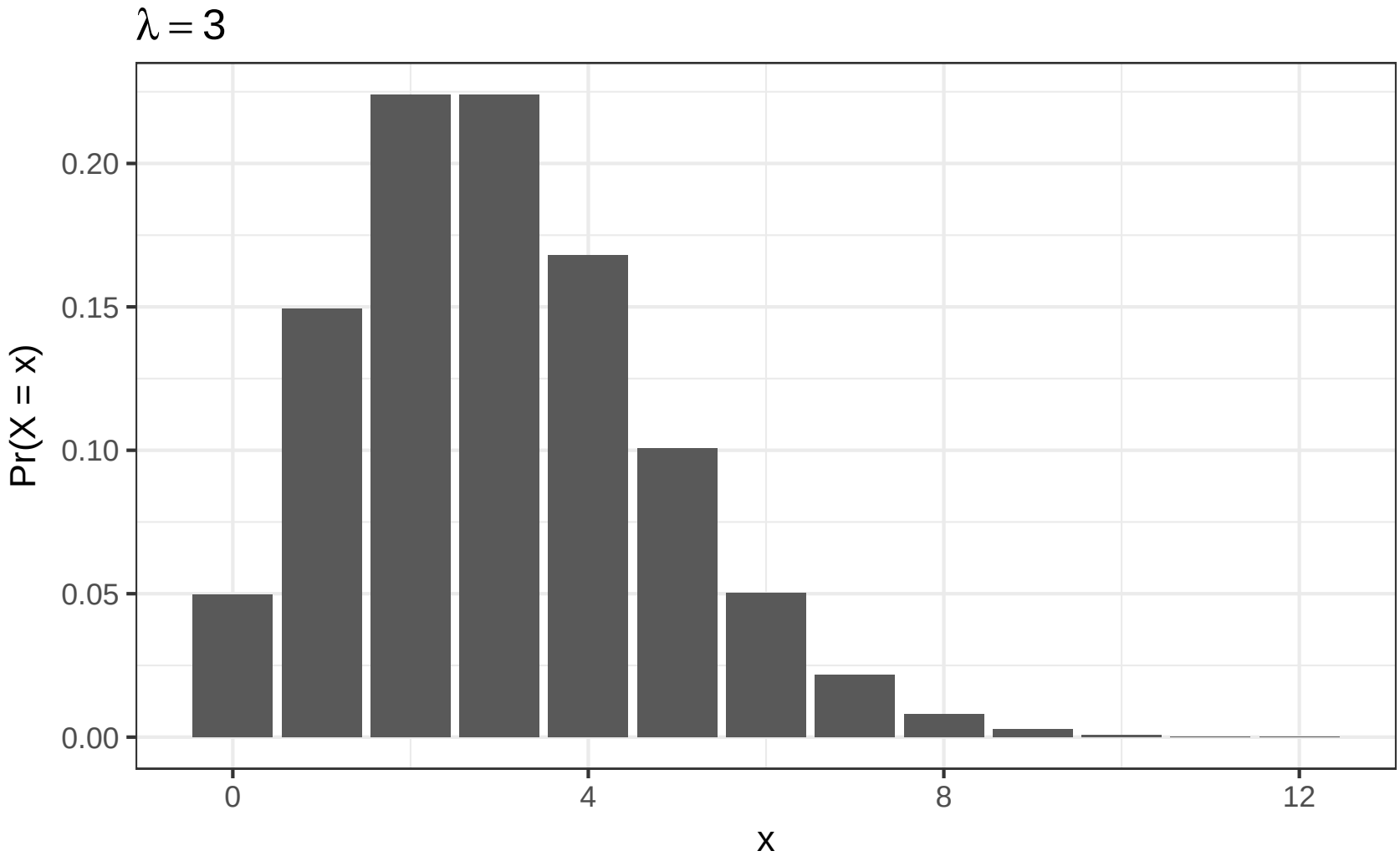
The most common starting point for analysing such data is by means of *Poisson regression*. This is a GLM with a Poisson probability distribution. This distribution is

$$\Pr(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!},$$

$y = 0, 1, 2, \ldots$. The parameter of interest is λ, which is interpreted as the *rate* of occurrence of events in a given unit of time. Because a rate must be non-negative, the most common link function is the exponential function

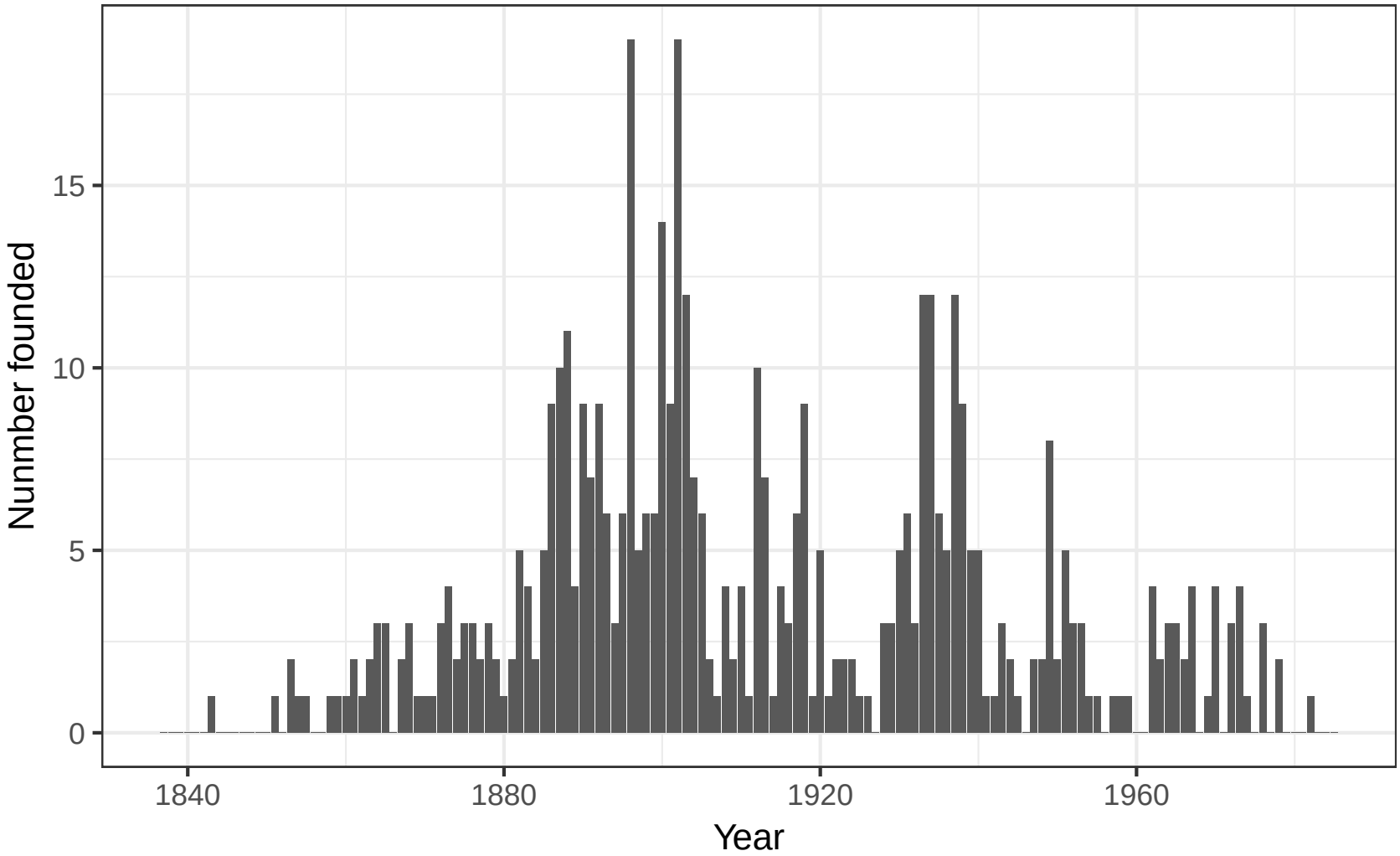
$$\begin{aligned} \eta(x) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}; \\ \lambda_i &= \exp(\eta(x)). \end{aligned}$$

Poisson distribution



Example: labour unions

Number of labour unions founded each year in the United States, 1837–1985.

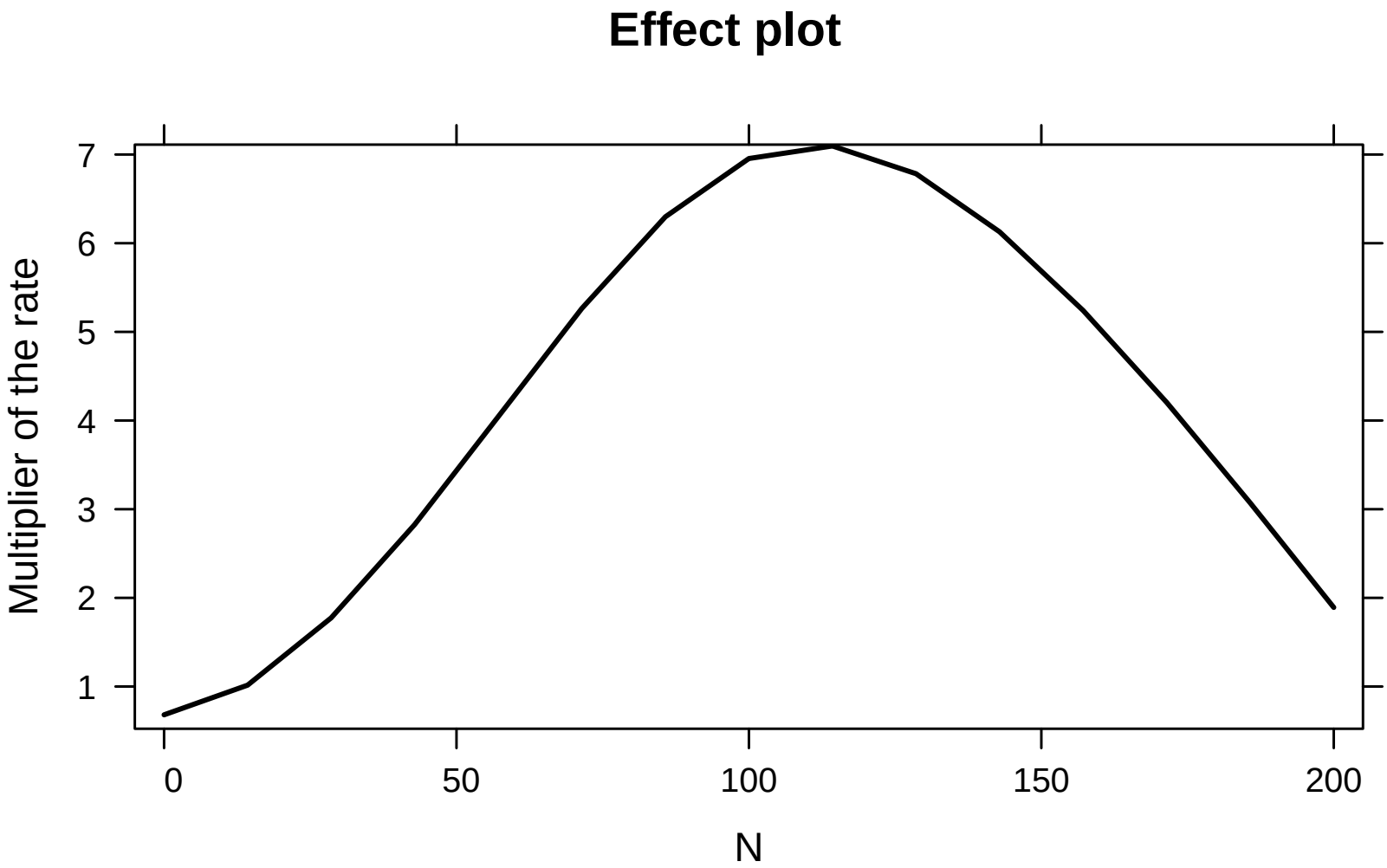


Poisson regression results

```
Call:
glm(formula = FND ~ poly(N, 2, raw = TRUE) + LAGF + AFL + NEWDEAL +
    TAFTH + AFLCIO + DEP, family = poisson, data = union)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.787  -1.229  -0.220   0.678   3.462

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -3.65e-01   1.68e-01   -2.18  0.02941
poly(N, 2, raw = TRUE)1  4.14e-02   4.23e-03    9.79  < 2e-16
poly(N, 2, raw = TRUE)2 -1.81e-04   2.08e-05   -8.72  < 2e-16
LAGF                     4.62e-02   1.19e-02    3.88  0.00011
AFL                     -2.30e-01   2.47e-01   -0.93  0.35345
NEWDEAL                  6.96e-01   1.92e-01    3.62  0.00029
TAFTH                    1.11e-01   2.51e-01    0.44  0.65721
AFLCIO                  -1.62e+00   2.93e-01   -5.53  3.3e-08
DEP                      8.45e-02   1.15e-01    0.73  0.46288

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 538.89  on 148  degrees of freedom
Residual deviance: 249.36  on 140  degrees of freedom
AIC: 616

Number of Fisher Scoring iterations: 5
```

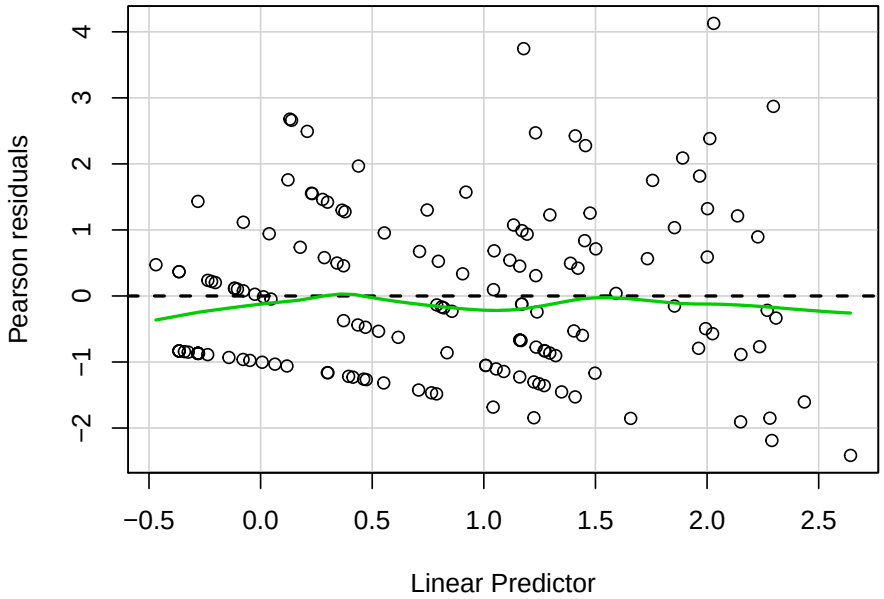
The results are reasonably straightforward to interpret. Each parameter and variable pair constitute a *multiplier* of the rate. So, for example, when there had been 10 foundings the previous year, the rate in the current year would be multiplied by $e^{0.462} = 1.587$.

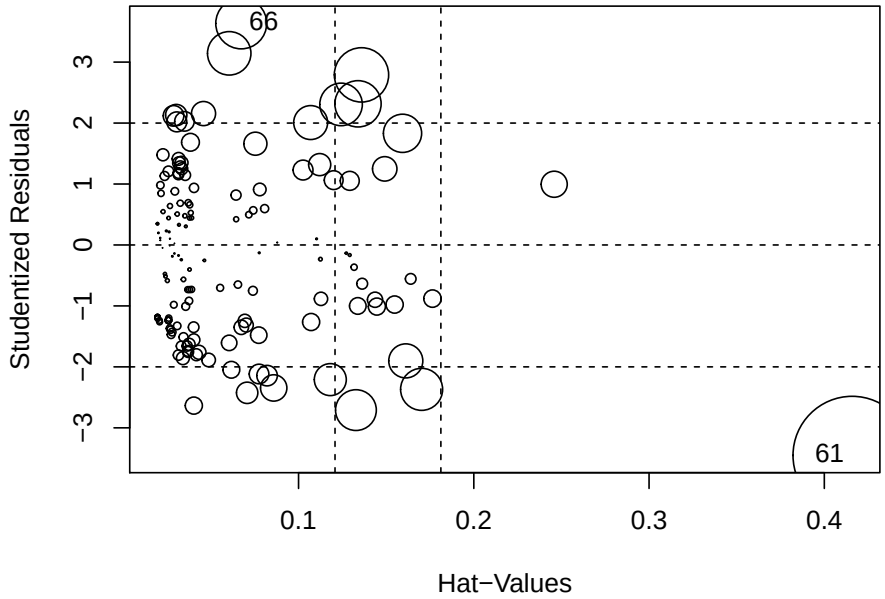
Effect plot



Residuals, outliers, etc.

As with linear and logistic regressions, it is possible to plot residuals and do various outlier tests.





	StudRes	Hat	CookD
61	-3.45	0.416	0.790
66	3.63	0.067	0.146

Negative binomial regression

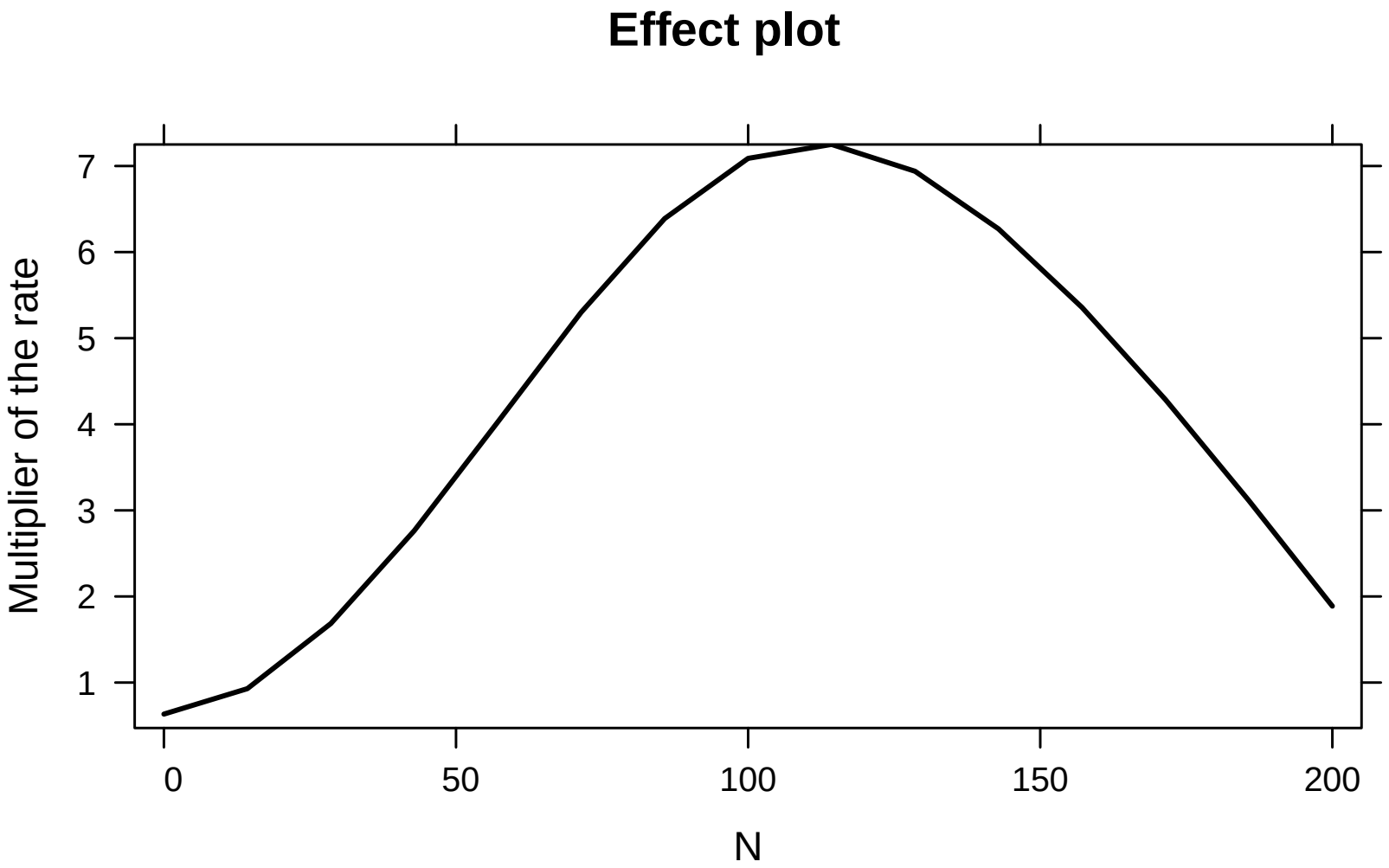
Overdispersion

The Poisson model is based on quite stringent assumptions. In our example, for instance, we are implicitly assuming that the rate of occurrence of foundings is constant during each year. However, given that our model asserts that the rate depends on N (which varies during the year), that seems implausible. If it is not in fact true, we in effect have additional sources of variation in the rate that are unmodelled, leading to **overdispersion**. The main problem caused by overdispersion is that estimated standard errors will be biased, usually downwards. That is, we are at risk of rejecting the null hypothesis even when it is true. We can solve this problem quite easily by using a method known as **negative binomial regression** as it has a negative binomial probability distribution. The link function is the same as in the case of Poisson regression. You can think of negative binomial regression as being Poisson regression with an additional parameter to model the overdispersion.

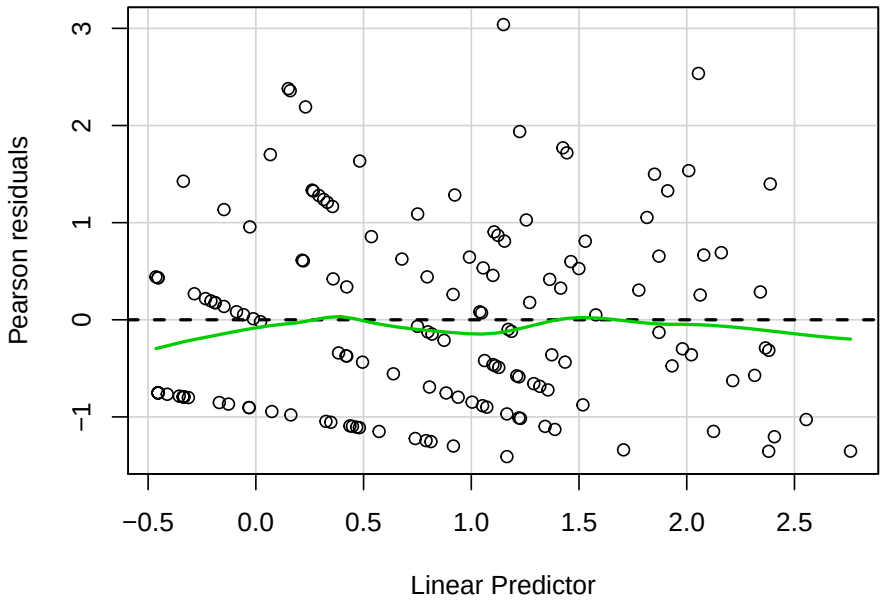
Negative binomial regression example

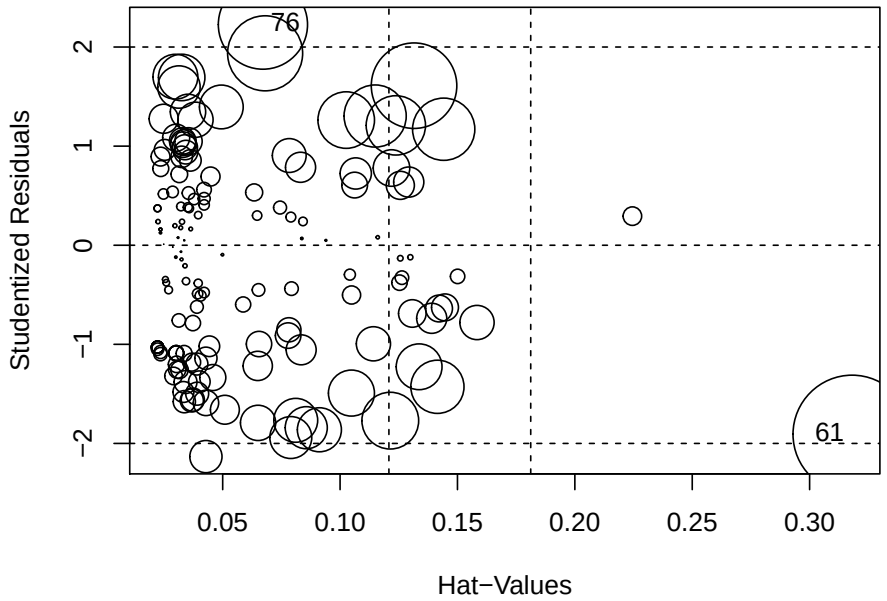
```
Call:
MASS::glm.nb(formula = FND ~ poly(N, 2, raw = TRUE) + LAGF +
    AFL + NEWDEAL + TAFTH + AFLCIO + DEP, data = union, init.theta = 5.200623228,
    link = log)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.235  -1.117  -0.151   0.553   2.208

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.54e-01   1.98e-01   -2.29   0.0220
poly(N, 2, raw = TRUE)1  4.28e-02   5.68e-03    7.54  4.6e-14
poly(N, 2, raw = TRUE)2 -1.87e-04   2.85e-05   -6.55  5.8e-11
LAGF                     5.57e-02   1.80e-02    3.10   0.0019
AFL                     -2.97e-01   3.77e-01   -0.79   0.4308
NEWDEAL                  6.51e-01   2.70e-01    2.41   0.0160
TAFTH                    2.19e-01   3.28e-01    0.67   0.5049
AFLCIO                  -1.60e+00   3.71e-01   -4.31  1.6e-05
DEP                      1.17e-01   1.54e-01    0.76   0.4465

(Dispersion parameter for Negative Binomial(5.2) family taken to be 1)

    Null deviance: 341.22  on 148  degrees of freedom
Residual deviance: 160.59  on 140  degrees of freedom
AIC: 591.3

Number of Fisher Scoring iterations: 1

              Theta:  5.20
          Std. Err.:  1.64

 2 x log-likelihood:  -571.27
```

Effect plot



Residuals, outliers, etc.





	StudRes	Hat	CookD
61	-1.91	0.318	0.139
76	2.23	0.067	0.079

Event history analysis

Basic principles

Analysis of time to events

Methods known as *event history analysis*, *survival analysis*, or *duration analysis* are used when we are interested in the length of time until an event occurs, known as an **episode** or **spell**. We generally need at least two pieces of information: how long the spell lasts and how it ended. We may also need to know the actual (calendar) time at which it started and ended, rather than just the duration. Generally, some spells will end with the event of interest and some will not. It is possible for there to be multiple possible outcomes, but we will not deal with these more complex cases.

An example would be the lifespan of organizations. If we know when they are founded and when they fail, then we can calculate the duration until they fail. In addition, most likely there will be some that have not (yet) failed at the time of our observation. These are called **censored** cases, and their existence is one of the main reasons we have to use special methods for the analysis of this type of data. We can't just ignore censored cases, but on the other hand we can't treat them as being the same as cases that have actually ended; both of these would introduce bias into regression parameter estimates.

Basic concepts

Hazard rate The instantaneous rate at which events occur:

$$r(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Survivor function. The probability that a unit doesn't experience an event before time t.

$$G(t) = 1 - F(t) = \Pr(T \geq t).$$

From this, we can define the hazard rate as

$$r(t) = \frac{f(t)}{G(t)}.$$

Survival data in R

The most commonly used package is `survival`, but I also use the package `eha`. Example data: duration of jobs.

```
rr.s <- Surv(rrdat1$dur, rrdat1$des)
```

The first variable is the duration and the second is the indicator of how the episode ended (0/1 or TRUE/FALSE). This will be the outcome variable in our regression analyses.

Example data: job durations

- `id` Identification number of subject
- `noj` Serial number of the job episode
- `ts` Starting time of the job episode
- `tf` Ending time of the job episode
- `sex` Sex
- `ti` Date of interview (CMC: Months since Dec 1899, ie, Jan 1900 = 1)
- `tb` Date of birth (CMC)
- `te` Date of entry into the labour market (CMC)
- `tmar` Date of marriage (CMC) [0 if not married]
- `pres` Prestige score of current job, i.e. of job episode in current record of data file
- `presn` Prestige score of the next job (if missing: -1)
- `edu` Highest educational attainment before entry into labour market
- `coho` Birth cohort (1: before 1940; 2: 1940–1949; 3: 1950 onwards)
- `lfx ts - te`
- `des tf != ti` (if spell ends on date of interview, spell is censored)
- `dur tf - ts + 1`

Kaplan-Meier estimator

There are q points in time at which at least one event occurs. There are I intervals between these q points. Then the Kaplan-Meier (or product limit) estimate of the survivor function is

$$\widehat{G}(t) = \prod_{l:\tau_l < t} \left(1 - \frac{E_l}{R_l}\right),$$

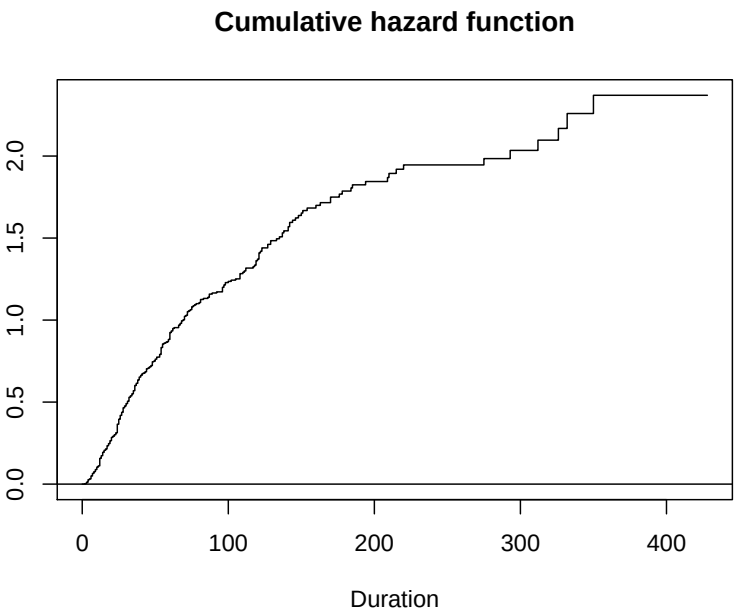
where E_l is the number of events in interval I and R_l is the risk set in the same interval. From this we can also calculate the cumulative hazard:

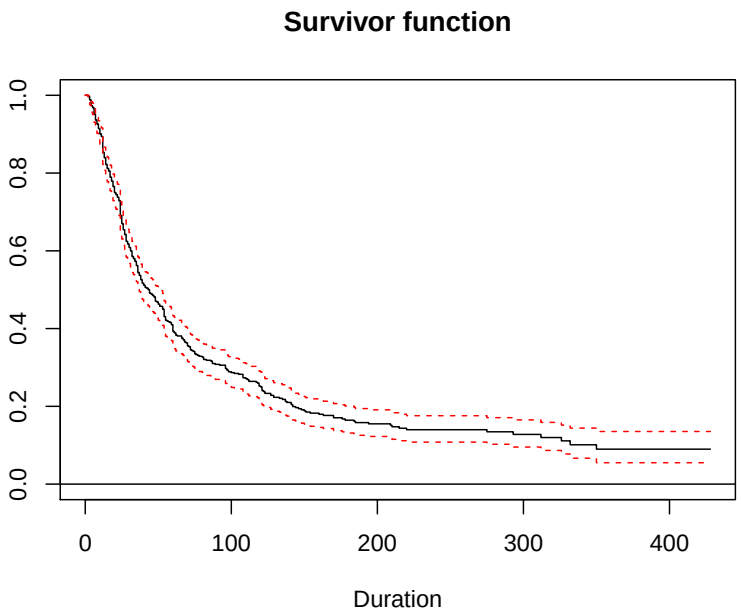
$$\widehat{H}(t) = -\log\left(\widehat{G}(t)\right).$$

Example

```
Call: survfit(formula = rr.s ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    600       0    1.000 0.00000        1.000        1.000
    2    600       2    0.997 0.00235        0.992        1.000
    3    597       5    0.988 0.00439        0.980        0.997
    4    590       9    0.973 0.00660        0.960        0.986
    5    581       3    0.968 0.00717        0.954        0.982
    6    577      10    0.951 0.00880        0.934        0.969
    7    567       9    0.936 0.00999        0.917        0.956
    8    557       6    0.926 0.01070        0.906        0.947
    9    548       7    0.914 0.01146        0.892        0.937
   10    540       8    0.901 0.01225        0.877        0.925
   11    528       4    0.894 0.01262        0.870        0.919
   12    524      24    0.853 0.01455        0.825        0.882
   13    499       8    0.839 0.01510        0.810        0.870
   14    488      10    0.822 0.01574        0.792        0.854
   15    477       6    0.812 0.01610        0.781        0.844
   16    471       4    0.805 0.01633        0.774        0.838
   17    467       9    0.789 0.01681        0.757        0.823
   18    458       6    0.779 0.01711        0.746        0.813
   19    452       8    0.765 0.01749        0.732        0.800
   20    443       9    0.750 0.01789        0.716        0.786
```

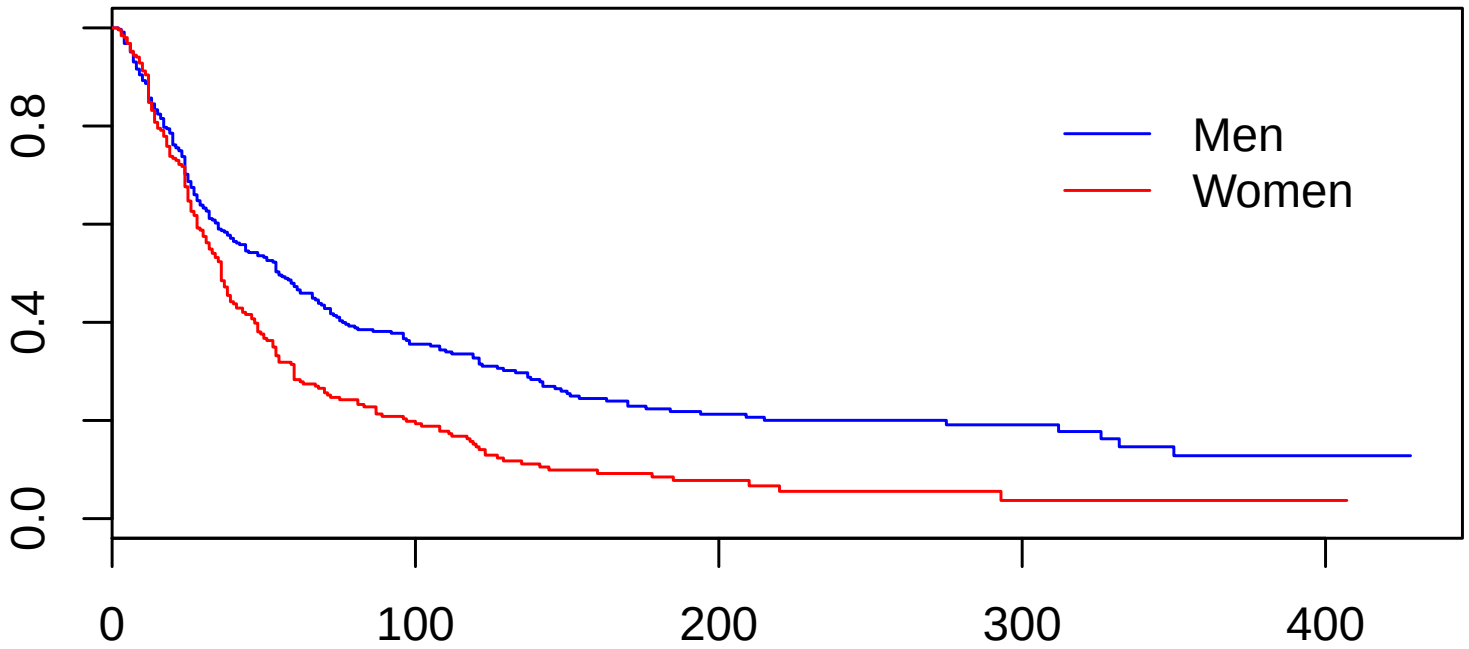
Estimates of G and H





Stratified survivor functions

```
Call: survfit(formula = rr.s ~ sex, data = rrdat1)

             n events median 0.95LCL 0.95UCL
sex=man    348    245     55      44      68
sex=woman  252    213     36      32      41
```



Continuous time models

Parametric rate models

Parametric rate models are used for events that occur in continuous time and where we are interested in finding out something about the nature of duration dependence on the rate. Define $c_i = 0$ for episodes that end in an event and $c_i = 1$ for those that are censored. The likelihood can then be written:

$$\mathcal{L} = \prod_{i=1}^{n} f(t_i)^{1-c_i} G(t_i)_i^c.$$

This is how we use the information about censored cases. We know that an observation that is censored at time t survived until at least time t. The probability of that is just the survivor function, $G(t)$. For events that occurred at time t, we use the probability density function, $f(t)$.

The basic model is

$$r(t) = r_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k).$$

$r_0(t)$ is called the *baseline hazard rate*.

Exponential distribution

The simplest distribution that we can use is the exponential distribution.

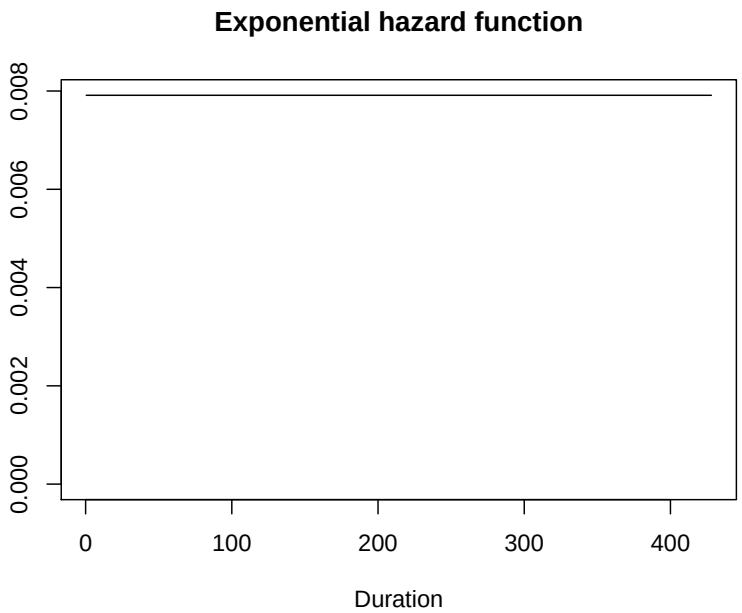
$$
\begin{aligned}
r(t) &= r_0 \exp(\beta_1 X_1 + \beta_2 X_2 + \dots) \\
G(t) &= \exp(-rt) \\
f(t) &= r \exp(-rt)
\end{aligned}
$$

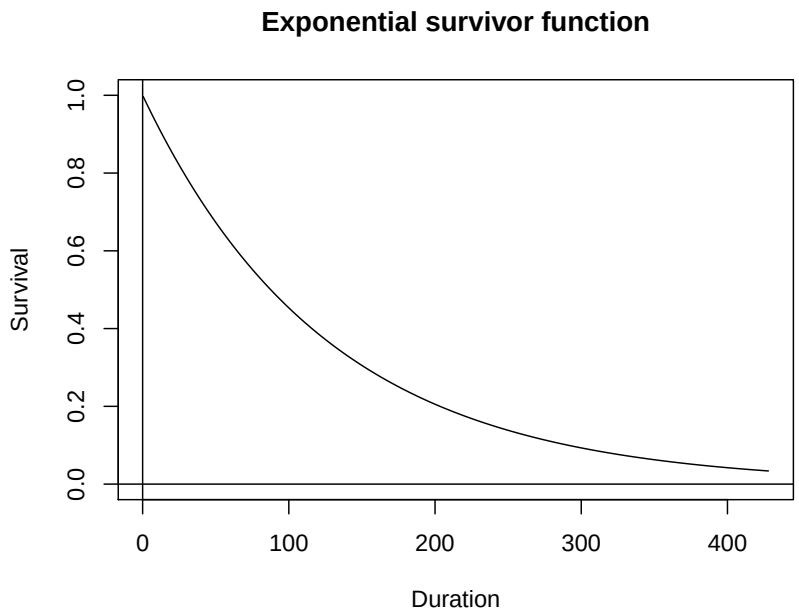
In other words, the baseline hazard rate is a constant. If events occur during some interval with an exponential distribution at a rate r, then a count of events during the same interval will have a Poisson distribution with mean $1/r$. Notice that this is the only model in which there is no time-dependence in the hazard rate.

Example

```
Call:
phreg(formula = rr.s ~ edu + coho + lfx + pnoj + pres, data = rrdat1,
    shape = 1, center = TRUE)

Covariate           W.mean      Coef Exp(Coef)  se(Coef)    Wald p
edu                 11.097     0.077     1.080     0.025     0.002
coho
           coho1     0.543     0         1           (reference)
           coho2     0.256     0.608     1.837     0.114     0.000
           coho3     0.201     0.611     1.842     0.119     0.000
lfx                 74.980    -0.003     0.997     0.001     0.001
pnoj                 1.406     0.060     1.061     0.044     0.177
pres                39.103    -0.028     0.972     0.006     0.000

log(scale)                     4.489                0.280     0.000

 Shape is fixed at  1

Events                    458
Total time at risk         40782
Max. log. likelihood      -2466
LR test statistic         7.24
Degrees of freedom        6
Overall p-value           0.298852
```

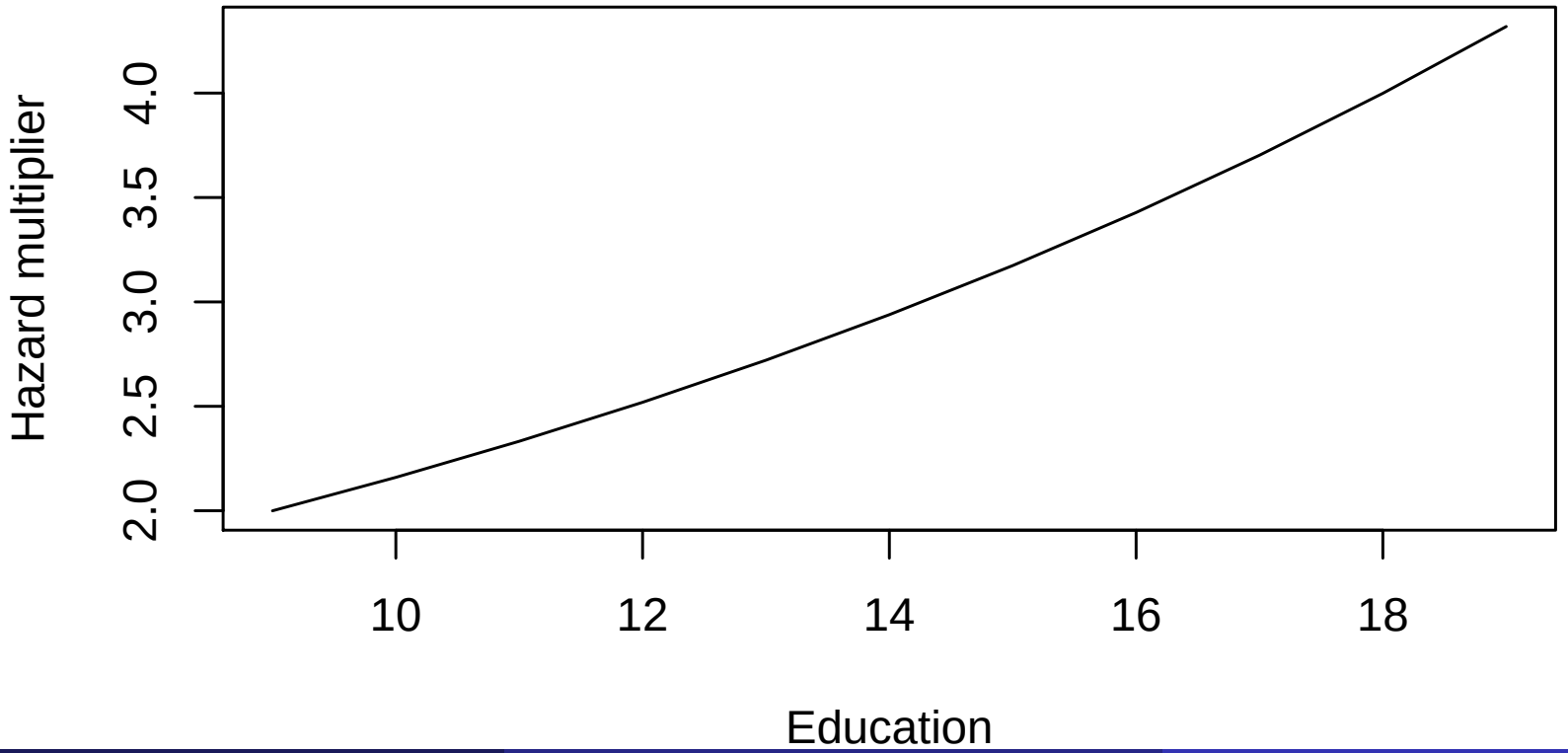
Estimated survivor and hazard functions





Interpretation

The baseline hazard rate ($r_0(t)$) is multiplied by the estimated effects. For example, the variable edu has a minimum value of 9 and a maximum of 19. At its minimum, the multiplier of the baseline hazard is $\exp(0.077 \times 9) = 2.00$, while at the maximum the multiplier is $\exp(0.077 \times 19) = 4.32$. The multiplier can be plotted:



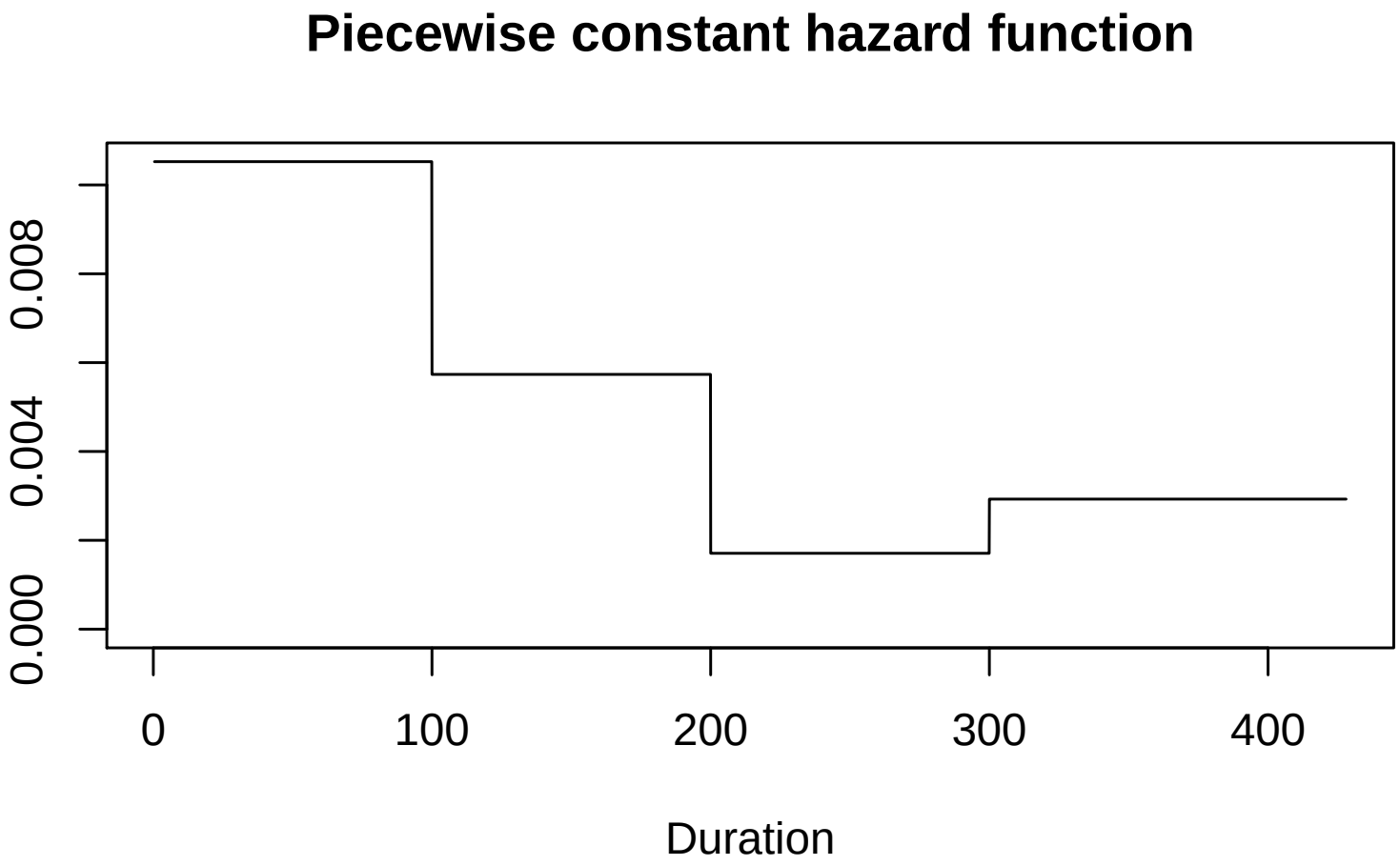
Piece-wise exponential

A very flexible alternative is to specify durations that will have the same rate, but allow the rate to differ across these periods. Suppose we hypothesize that the hazard rate differs for survival times less than 100 months, between 100 and 200 months, 200 and 300 months, and over 300 months, but is constant within these periods.

Example

```
Call:
phreg(formula = rr.s ~ edu + coho + lfx + pnoj + pres, data = rrdat1,
    dist = "pch", cuts = c(100, 200, 300))

Covariate           W.mean      Coef Exp(Coef)  se(Coef)    Wald p
edu                 11.097     0.068     1.071     0.025     0.006
coho
           coho1     0.543     0         1           (reference)
           coho2     0.256     0.439     1.550     0.115     0.000
           coho3     0.201     0.334     1.397     0.122     0.006
lfx                 74.980    -0.004     0.996     0.001     0.000
pnoj                 1.406     0.069     1.071     0.044     0.118
pres                39.103    -0.026     0.974     0.005     0.000


Events                    458
Total time at risk         40782
Max. log. likelihood      -2438.9
LR test statistic         77.92
Degrees of freedom        6
Overall p-value           9.54792e-15
```

Piecewise constant hazard



Goodness of fit test

The exponential and piecewise exponential are nested models, so can be tested against each other using a simple likelihood ratio test.

$$LR = -2(L_0 - L_1)$$

where L_0 is the maximum log likelihood from the simpler model, and L_1 is the equivalent for the more complex model. The test statistic has a chi-square distribution with degrees of freedom equal to the number of extra parameters in the second model.

$$LR = -2 \times (-2466 - -2439) = 54$$

with 3 degrees of freedom, which is highly significant, so we can conclude that the piecewise model fits better than the simple exponential model.

Weibull hazard rate

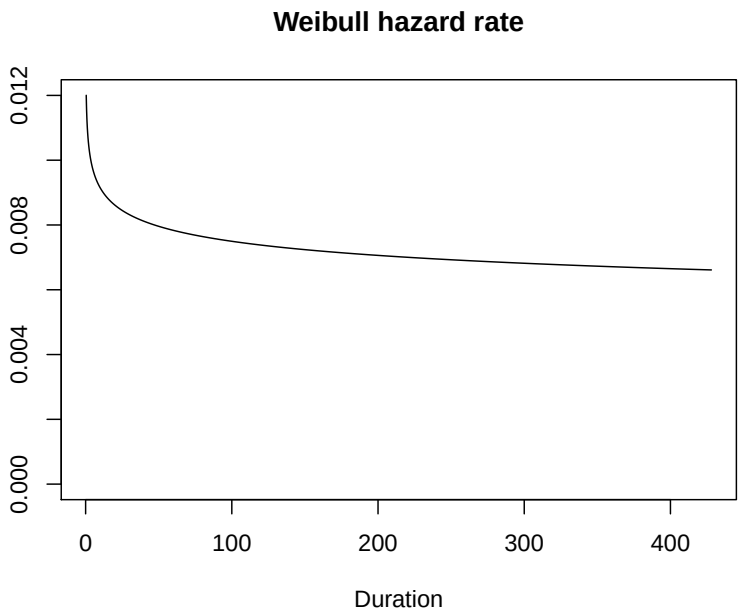
$$r(t) = \frac{p}{\lambda}\left(\frac{t}{\lambda}\right)^{(p-1)} \exp(\beta_1 X_1 + \beta_2 X_2 + \dots)$$

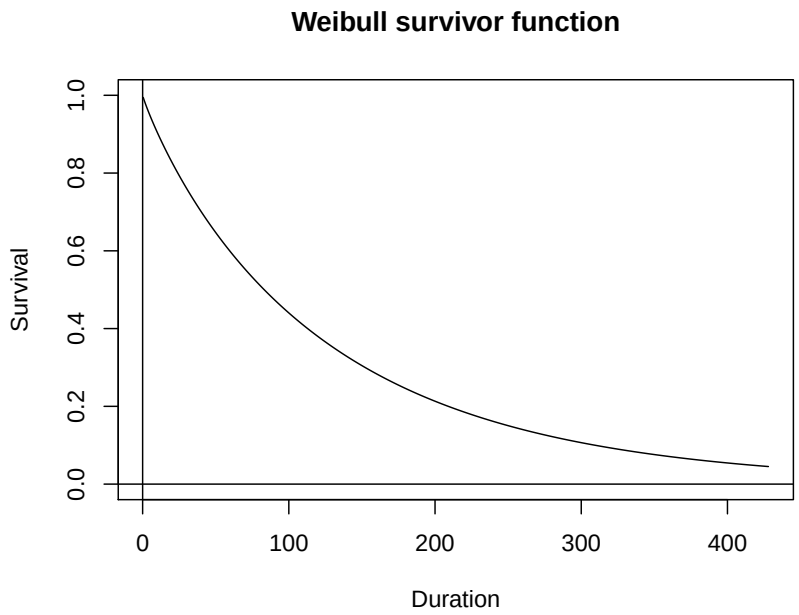
The exponential model is obtained when $p = 1$, so it is straightforward to do a hypothesis test of Weibull against exponential. In the exponential model, $r = 1/\lambda$.

Example

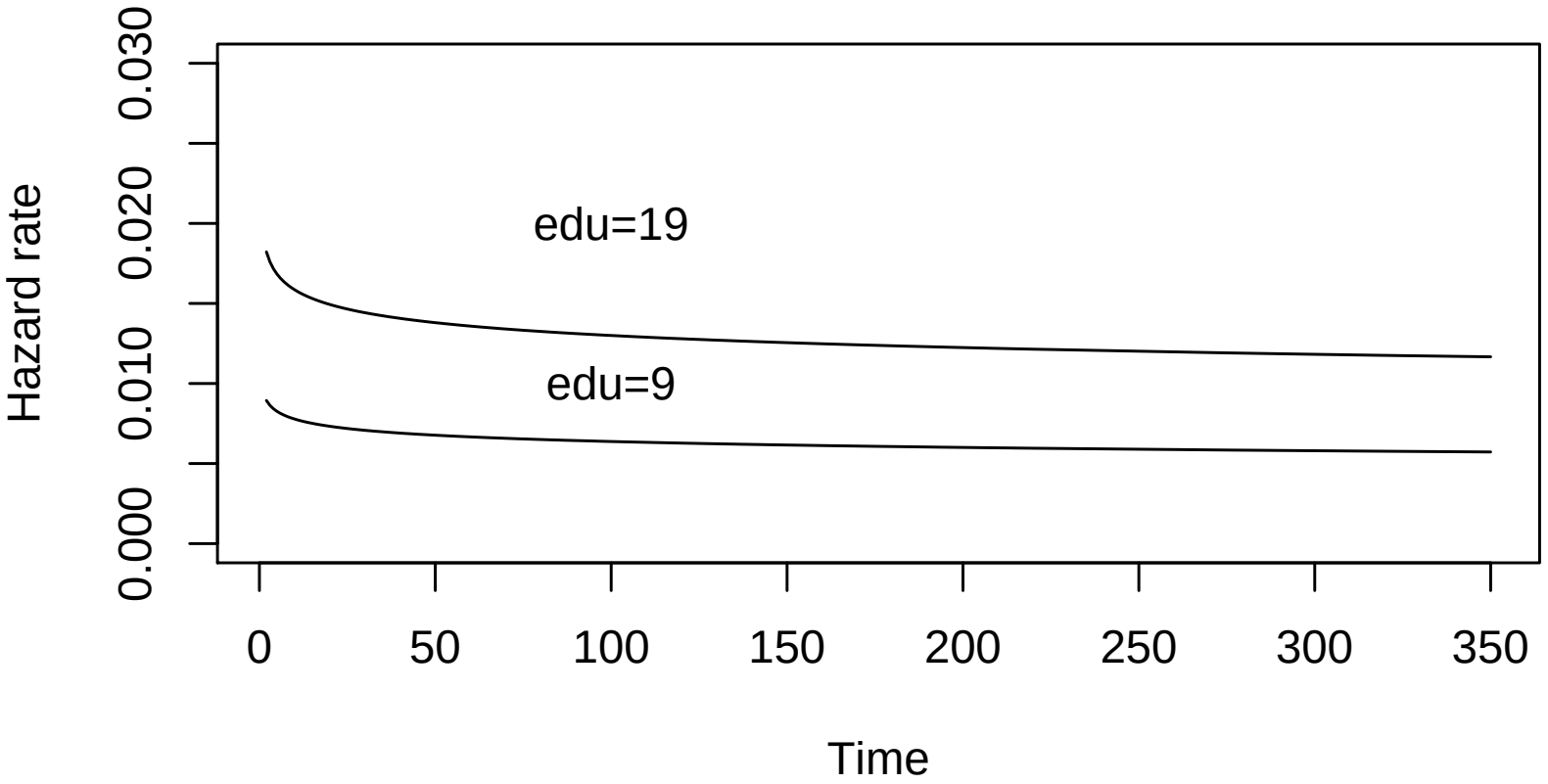
```
Call:
phreg(formula = rr.s ~ edu + coho + lfx + pnoj + pres, data = rrdat1,
    x = TRUE)

Covariate           W.mean      Coef Exp(Coef)  se(Coef)    Wald p
edu                 11.097     0.071     1.074     0.025     0.004
coho
           coho1     0.543     0         1           (reference)
           coho2     0.256     0.554     1.740     0.115     0.000
           coho3     0.201     0.528     1.695     0.123     0.000
lfx                 74.980    -0.003     0.997     0.001     0.000
pnoj                 1.406     0.058     1.060     0.044     0.187
pres                39.103    -0.027     0.974     0.006     0.000

log(scale)                     4.406               0.306     0.000
log(shape)                    -0.090               0.036     0.013

Events                     458
Total time at risk          40782
Max. log. likelihood      -2462.8
LR test statistic         84.28
Degrees of freedom        6
Overall p-value           4.44089e-16
```

Rate plots


Weibull hazard rate
0.000
0.004
0.008
0.012
0
100
200
300
400
Duration


Weibull survivor function
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
Duration

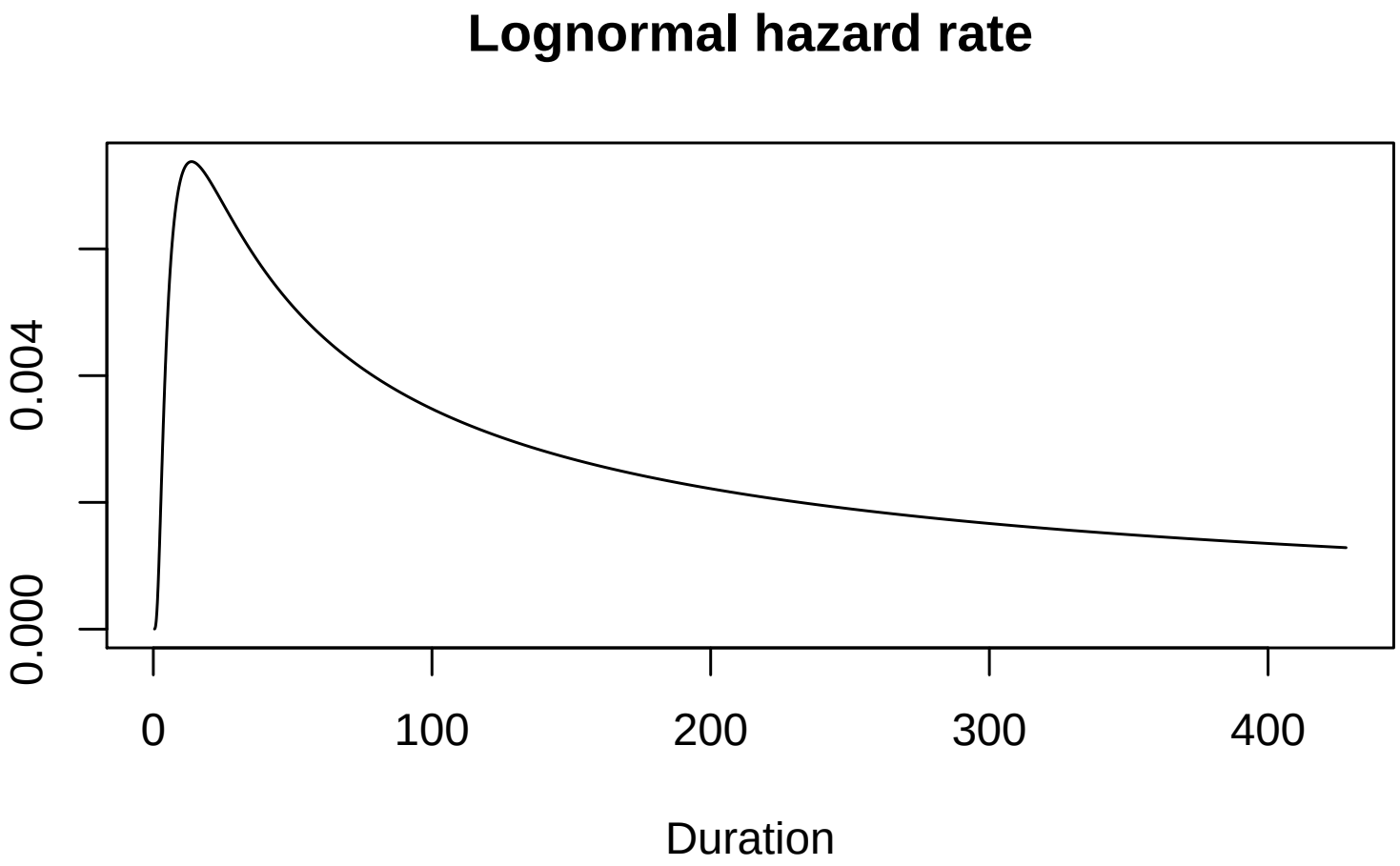
Predicted hazard



Lognormal distribution

```
Call:
phreg(formula = rr.s ~ edu + coho + lfx + pnoj + pres, data = rrdat1,
    dist = "lognormal")

Covariate           W.mean      Coef Exp(Coef)  se(Coef)    Wald p
(Intercept)                   -1.535               0.586     0.009
edu                11.097     0.068     1.070     0.025     0.006
coho
           coho1    0.543     0         1           (reference)
           coho2    0.256     0.448     1.566     0.115     0.000
           coho3    0.201     0.353     1.423     0.120     0.003
lfx                74.980    -0.004     0.996     0.001     0.000
pnoj                1.406     0.066     1.068     0.044     0.137
pres               39.103    -0.026     0.974     0.005     0.000

log(scale)                    2.425               0.346     0.000
log(shape)                    0.349               0.195     0.074

Events                    458
Total time at risk         40782
Max. log. likelihood      -2410.8
LR test statistic         77.22
Degrees of freedom        6
Overall p-value           1.34337e-14
```

Lognormal hazard



Cox regression

This model may be written

$$\log[r(t)] = \beta_0(t) + \beta_1 x_1(t) + \beta_2 x_2(t) + \cdots ,$$

When the covariates are constant over time, the ratio of the hazard rates for any pair of individuals will not depend on time. The partial likelihood estimator discards information about time, using only the order in which events occurred. This means some loss of efficiency, but it is typically very small. Also, we cannot obtain estimates of the dependence of the hazard on time. In many applications this doesn't matter; we are only interested in the effects of covariates.

Example

```
Call:
coxreg(formula = rr.s ~ edu + coho + lfx + pnoj + pres, data = rrdat1)

Covariate           Mean       Coef     Rel.Risk   S.E.     Wald p
edu                11.097     0.068     1.070     0.025     0.007
coho
           coho1    0.543     0         1 (reference)
           coho2    0.256     0.416     1.515     0.115     0.000
           coho3    0.201     0.309     1.362     0.122     0.011
lfx                74.980    -0.004     0.996     0.001     0.000
pnoj                1.406     0.069     1.071     0.044     0.118
pres               39.103    -0.027     0.974     0.006     0.000

Events                    458
Total time at risk         40782
Max. log. likelihood      -2542.2
LR test statistic         76.93
Degrees of freedom        6
Overall p-value           1.53211e-14
```

Discrete time

Discrete time models

Used when events can only occur at fixed, discrete time points. Sometimes also used when data can only be *measured* at discrete time points, typically whether an event occurred at some point during a year. If $P_i(t)$ is the probability of individual i experiencing an event at time t, we can use logistic regression:

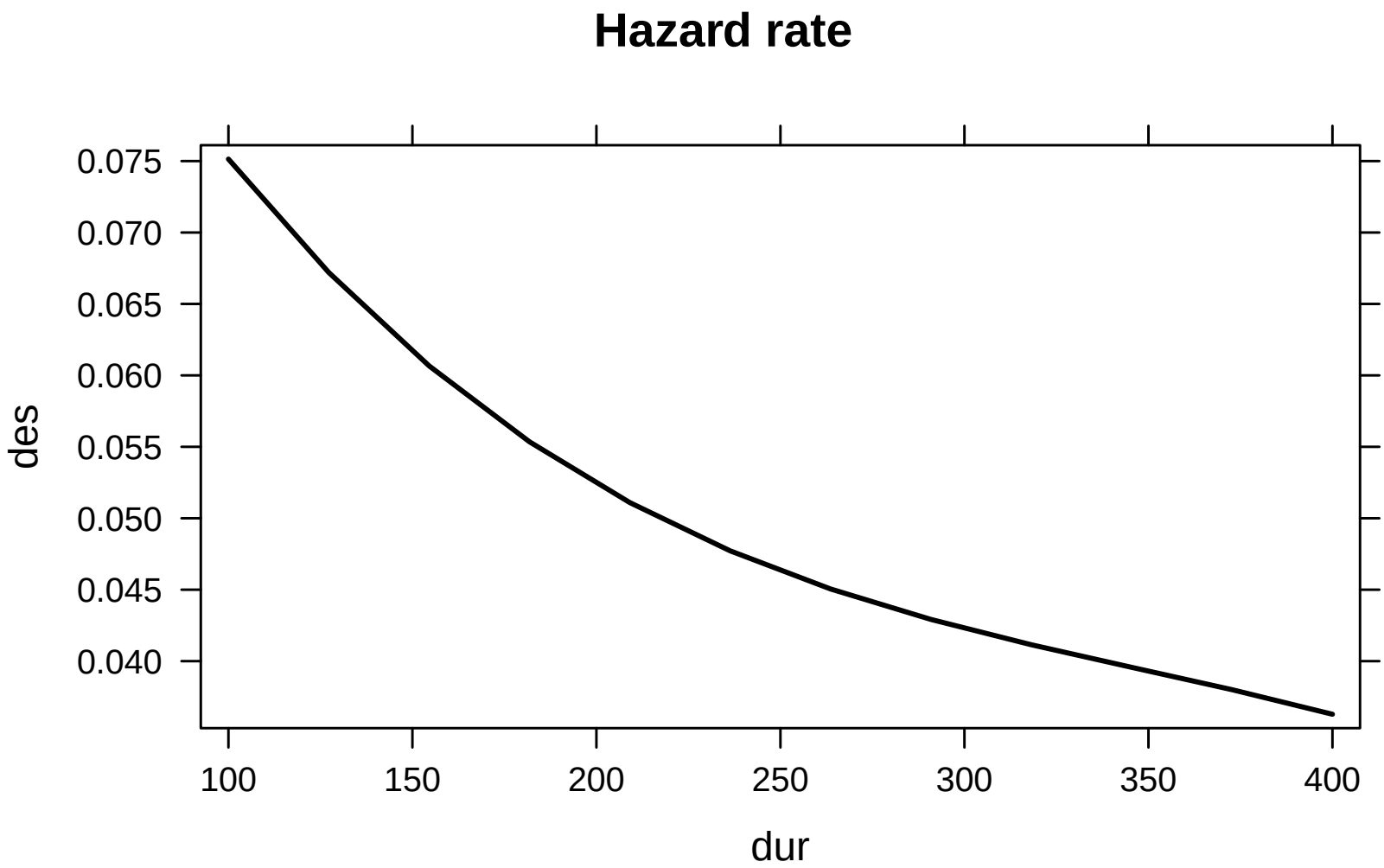
$$\log\left(\frac{P_i(t)}{1-P_i(t)}\right) = b_0(t) + b_1 x_1(t) + b_2 x_2(t) + \ldots.$$

If data are collected annually, then each unit has one observation per year until it experiences an event or is censored. $b_0(t)$ is the baseline hazard. In the example, I define it to be $\log(t)$.

Example

```
Call:
glm(formula = des ~ log(dur) + edu + coho + lfx + pnoj + pres,
    family = binomial, data = rrdat2)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.038  -0.564  -0.423  -0.298   2.684

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.475527   0.372167    1.28   0.2013
log(dur)    -0.554884   0.054545  -10.17  < 2e-16
edu          0.054748   0.027862    1.96   0.0494
cohocoho2    0.399447   0.125835    3.17   0.0015
cohocoho3    0.265145   0.133988    1.98   0.0478
lfx         -0.004225   0.000975   -4.33  1.5e-05
pnoj         0.051997   0.048169    1.08   0.2804
pres        -0.024409   0.005990   -4.08  4.6e-05

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2762.0  on 3667  degrees of freedom
Residual deviance: 2553.2  on 3660  degrees of freedom
AIC: 2569

Number of Fisher Scoring iterations: 5
```

Estimated hazard rate



Education effect