Week 4 Practical Session

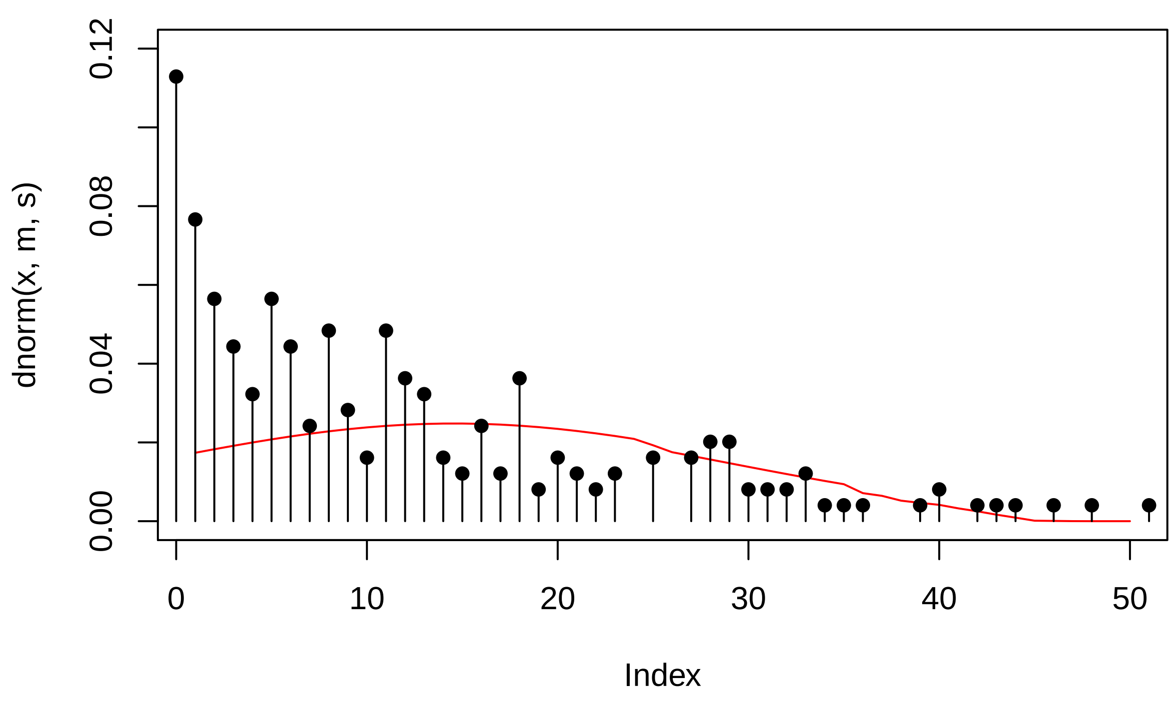
David Barron

Hilary Term 2018

Poisson regression

Poisson regression is used with count data. An example is data on interlocking directorates in 248 major Canadian firms in 1976. An "interlock" is created when two firms share one or more directors. Let's look at the data:

```
data(Ornstein)
f <- xtabs(~interlocks, Ornstein)
m <- mean(Ornstein$interlocks)
s <- sd(Ornstein$interlocks)
x <- as.numeric(names(f))
plot(dnorm(x, m, s), col = "red", type = "l", ylim = c(0, 0.12))
lines(x, f/sum(f), type = "h")
points(x, f/sum(f), pch = 16)
```



A variable like this could be analysed using linear regression, but it's not hard to see that it is a long way from being normally distributed. So, let's try poisson regression.

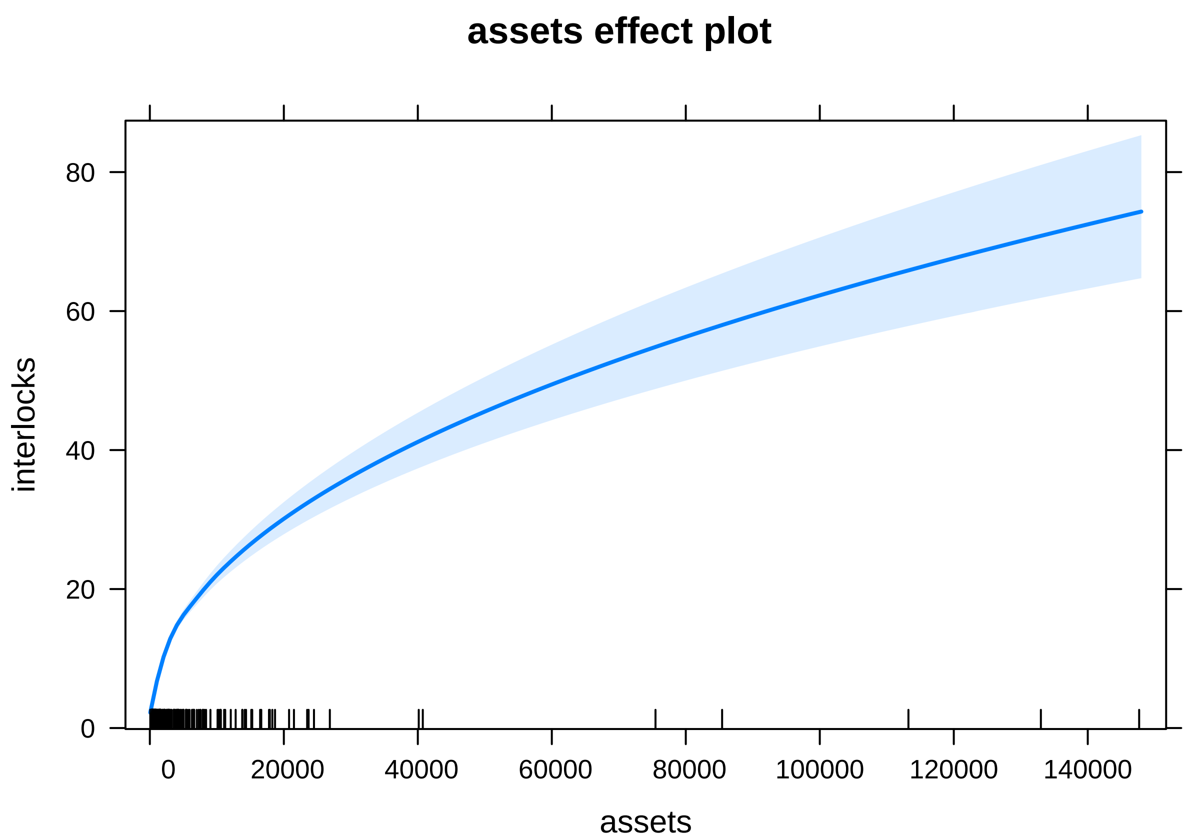
```
p1 <- glm(interlocks ~ log(assets) + nation + sector, family = poisson, data = Ornstein)
summary(p1)
```

```
Call:
glm(formula = interlocks ~ log(assets) + nation + sector, family = poisson,
    data = Ornstein)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.7111  -2.3159  -0.4595   1.2824   6.2849

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83938    0.13664  -6.143 8.09e-10 ***
log(assets)  0.45145    0.01698  26.585  < 2e-16 ***
nationOTH   -0.10699    0.07438  -1.438 0.150301
nationUK    -0.38722    0.08951  -4.326 1.52e-05 ***
nationUS    -0.77239    0.04963 -15.562  < 2e-16 ***
sectorBNK   -0.16651    0.09575  -1.739 0.082036 .
sectorCON   -0.48928    0.21320  -2.295 0.021736 *
sectorFIN   -0.11161    0.07571  -1.474 0.140457
sectorHLD   -0.01491    0.11924  -0.125 0.900508
sectorMAN    0.12187    0.07614   1.600 0.109489
sectorMER    0.06157    0.08670   0.710 0.477601
sectorMIN    0.24985    0.06888   3.627 0.000286 ***
sectorTRN    0.15181    0.07893   1.923 0.054453 .
sectorWOD    0.49825    0.07560   6.590 4.39e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3737.0  on 247  degrees of freedom
Residual deviance: 1547.1  on 234  degrees of freedom
AIC: 2473.1

Number of Fisher Scoring iterations: 5
```

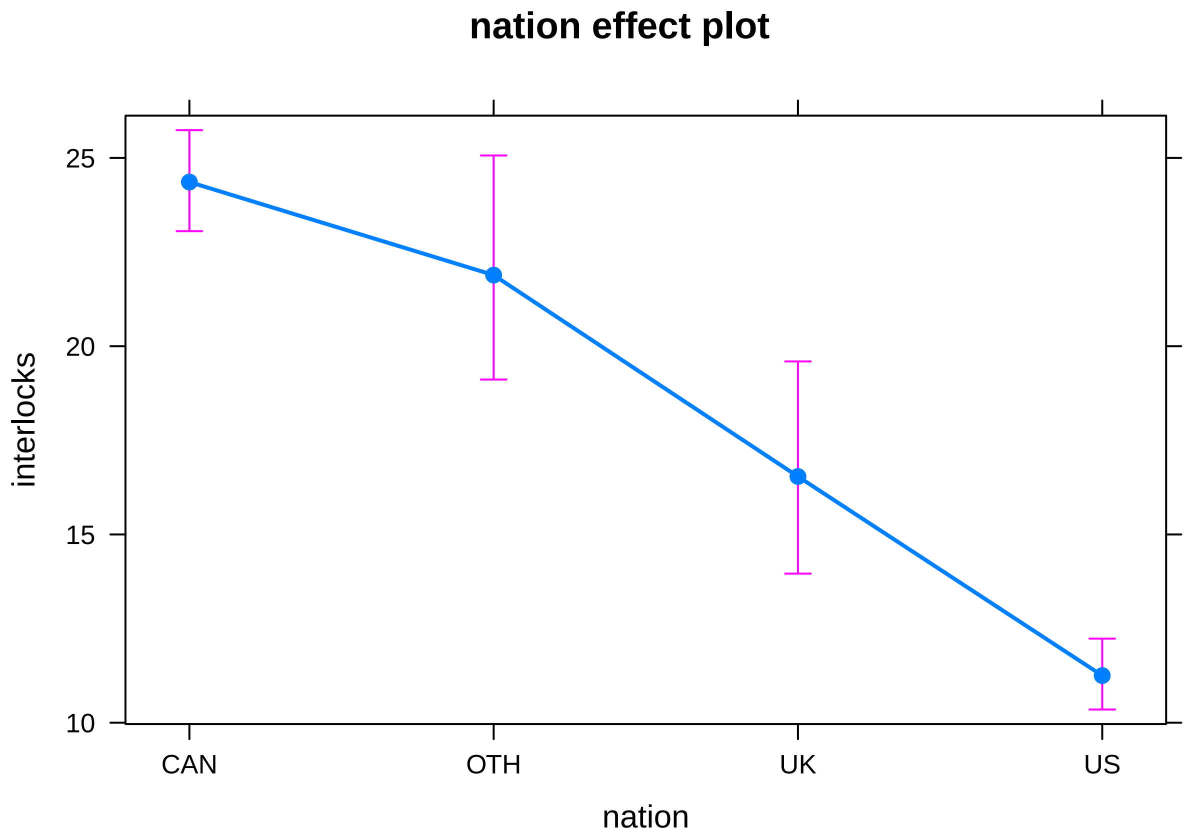
```r
Anova(p1)
```

```
Analysis of Deviance Table (Type II tests)

Response: interlocks
            LR Chisq Df Pr(>Chisq)
log(assets)   731.21  1  < 2.2e-16 ***
nation        276.04  3  < 2.2e-16 ***
sector        102.71  9  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

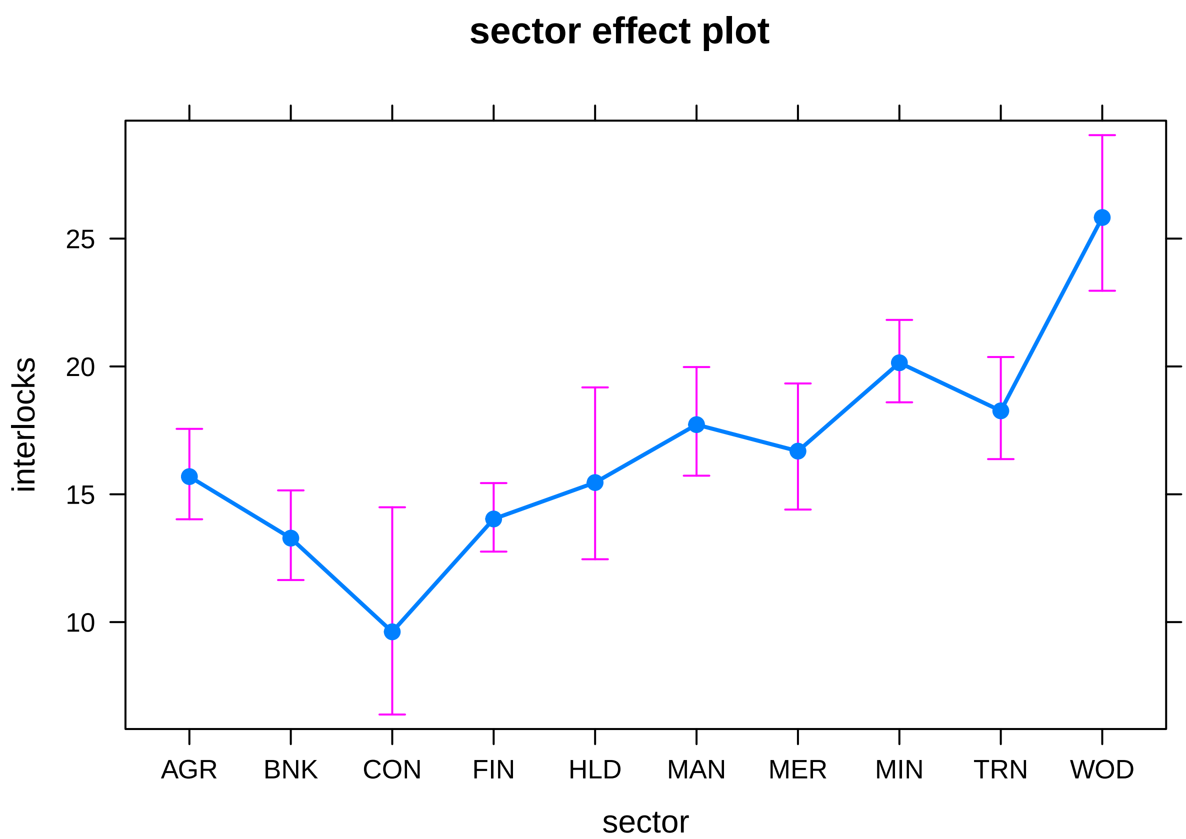
```r
plot(Effect("assets", p1, xlevels = list(assets = 50)), type = "response")
```




```
plot(Effect("nation", p1), type = "response")
```




```r
plot(Effect("sector", p1), type = "response")
```



Let's compare with negative binomial regression.

```
p2 <- glm.nb(interlocks ~ log(assets) + nation + sector, data = Ornstein)
summary(p2)
```

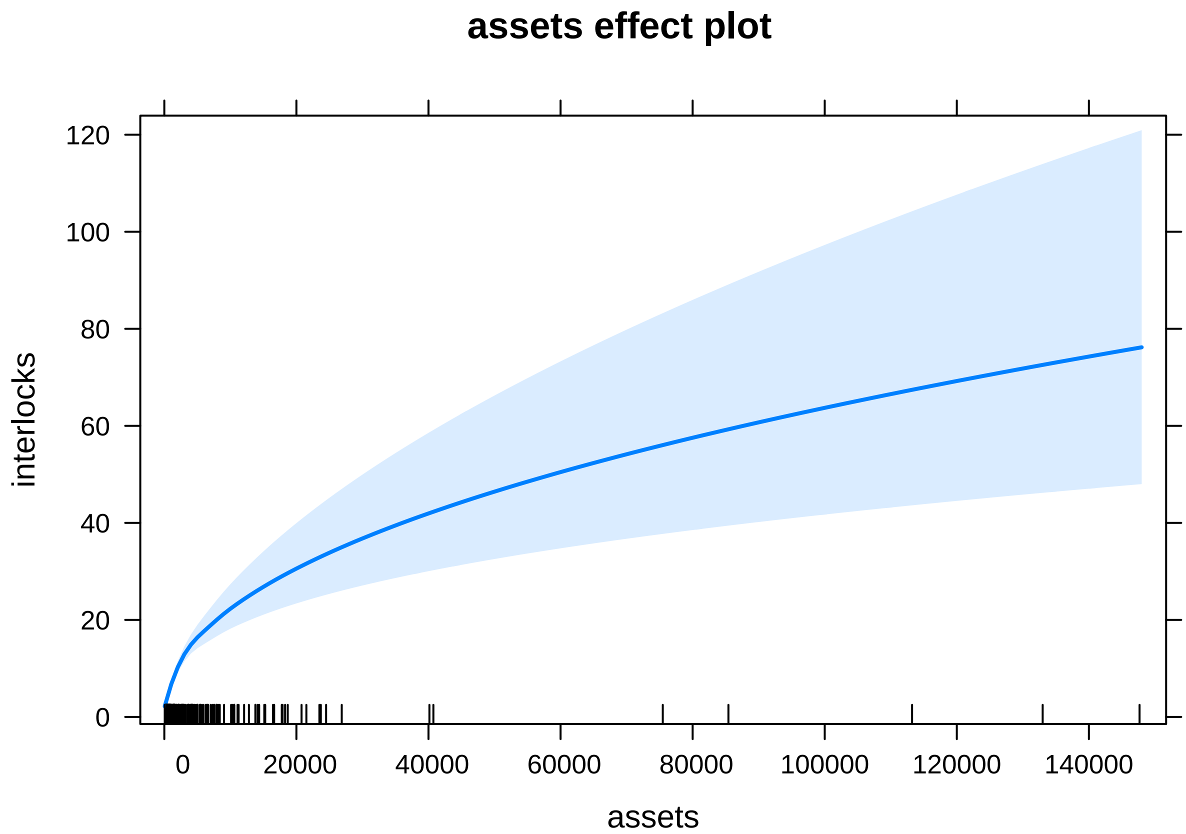
```
Call:
glm.nb(formula = interlocks ~ log(assets) + nation + sector,
    data = Ornstein, init.theta = 1.639034209, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8087  -0.9897  -0.1886   0.4301   2.4080

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.82535    0.37976  -2.173   0.0298 *
log(assets)  0.45618    0.05185   8.799  < 2e-16 ***
nationOTH   -0.10455    0.23004  -0.454   0.6495
nationUK    -0.38945    0.23575  -1.652   0.0985 .
nationUS    -0.78820    0.13201  -5.971 2.36e-09 ***
sectorBNK   -0.40846    0.37726  -1.083   0.2789
sectorCON   -0.75698    0.45711  -1.656   0.0977 .
sectorFIN   -0.10346    0.25181  -0.411   0.6812
sectorHLD   -0.21103    0.34982  -0.603   0.5463
sectorMAN    0.07677    0.18601   0.413   0.6798
sectorMER    0.07761    0.23246   0.334   0.7385
```

```
sectorMIN    0.23988    0.18837   1.273   0.2029
sectorTRN    0.10133    0.24752   0.409   0.6823
sectorWOD    0.39084    0.23253   1.681   0.0928 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.639) family taken to be 1)

    Null deviance: 521.58  on 247  degrees of freedom
Residual deviance: 296.52  on 234  degrees of freedom
AIC: 1675.3

Number of Fisher Scoring iterations: 1

              Theta:  1.639
          Std. Err.:  0.192

 2 x log-likelihood:  -1645.257
```

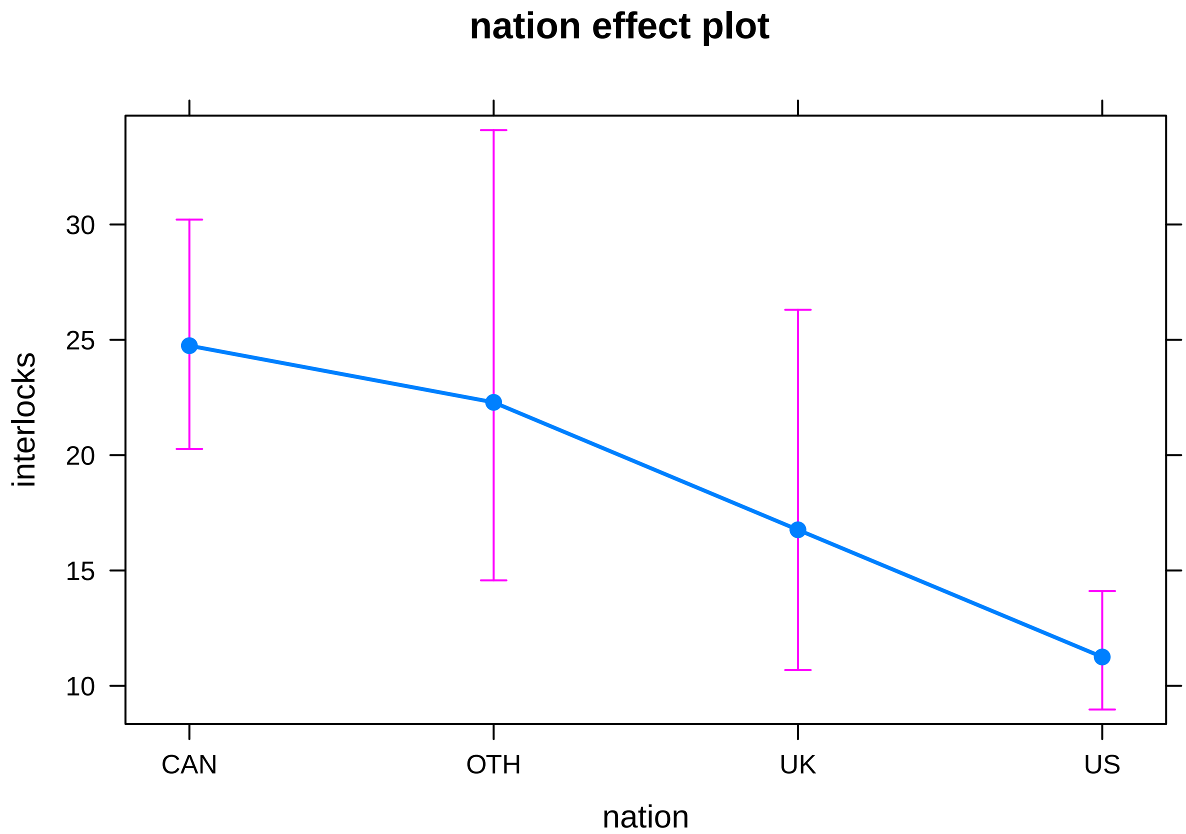
```r
Anova(p2)
```

```
Analysis of Deviance Table (Type II tests)

Response: interlocks
            LR Chisq Df Pr(>Chisq)
log(assets)   78.366  1  < 2.2e-16 ***
nation        38.030  3  2.786e-08 ***
sector        12.026  9     0.2118
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

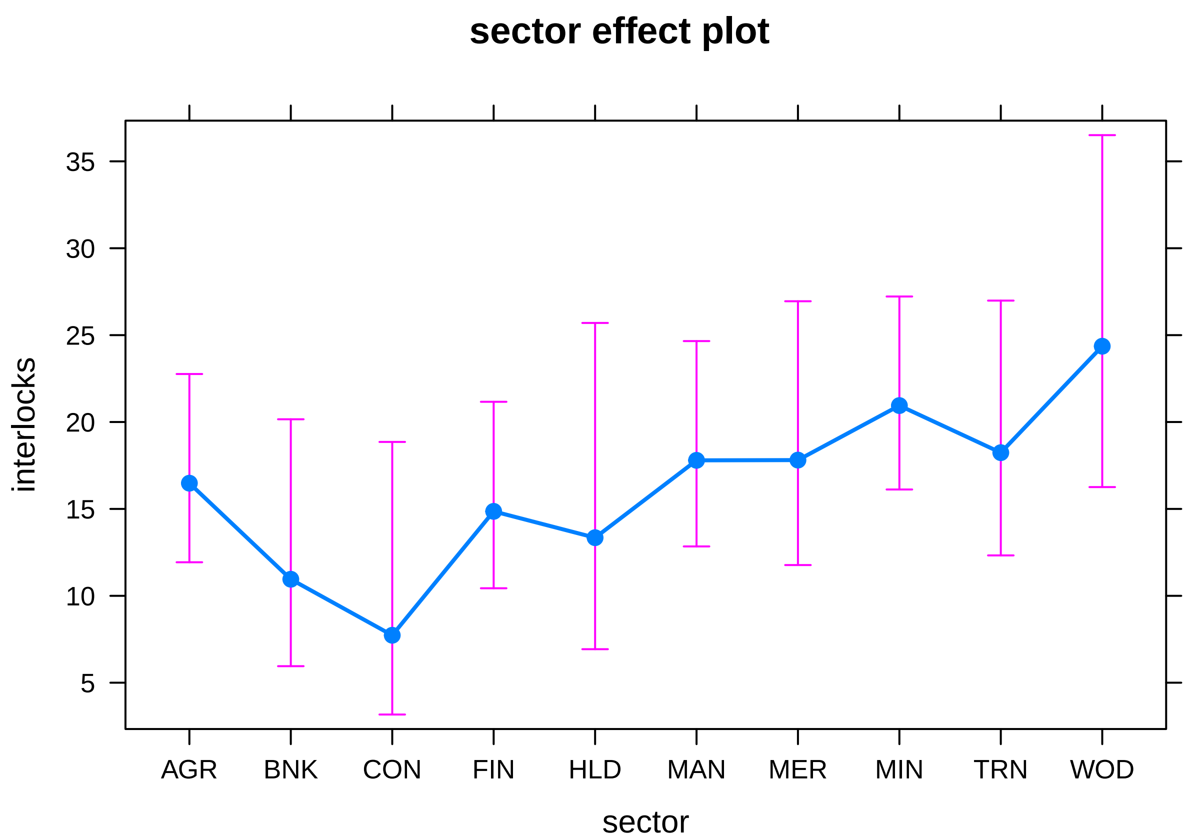
```r
plot(Effect("assets", p2, xlevels = list(assets = 50)), type = "response")
```




```
plot(Effect("nation", p2), type = "response")
```




```
plot(Effect("sector", p2), type = "response")
```



The negative binomial is clearly a better fit. This is very common in practice. The parameter estimates are very similar in the two sets of results, but notice that standard errors in the negative binomial results are larger. In fact `sector` no longer improves the fit of the model. This is common, and is one of the main reasons for preferring negative binomial regression; estimates of standard errors can be severly biased when there is significant overdispersion.

Event history analysis

Descriptive

We will use a dataset called `Rossi` in the `GlobalDeviance` package. These data are about recidivism in a group of 432 male prisoners, who were observed for a year after being released from prison. The variables are:

- `week`: week of first arrest after release, or censoring time;
- `arrest`: indicator, 1 if person arrested during perion of study, 0 otherwise;
- `fin`: indicator, 1 if person received financial support after release, 0 otherwise;
- `age`: at time of release;
- `race`: indicator, 1 = `black` or 0 = `other`;
- `wexp`: indicator, 1 if person had full-time work experience prior to prison, 0 otherwise;
- `mar`: indicator, 1 = `married` at time of release, 0 = `non married` otherwise;
- `paro`: indicator, 1 if person was released on parole, 0 otherwise;
- `educ`: level of education, in 6 categories;
- `emp1-emp52`: 1 if person employed in corresponding week, 0 otherwise.

```r
data(Rossi)
summary(survfit(Surv(Rossi$week, Rossi$arrest) ~ 1, data = Rossi))
```

```
Call: survfit(formula = Surv(Rossi$week, Rossi$arrest) ~ 1, data = Rossi)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    432       1    0.998 0.00231        0.993        1.000
    2    431       1    0.995 0.00327        0.989        1.000
    3    430       1    0.993 0.00400        0.985        1.000
    4    429       1    0.991 0.00461        0.982        1.000
    5    428       1    0.988 0.00515        0.978        0.999
    6    427       1    0.986 0.00563        0.975        0.997
    7    426       1    0.984 0.00607        0.972        0.996
    8    425       5    0.972 0.00791        0.957        0.988
    9    420       2    0.968 0.00852        0.951        0.984
   10    418       1    0.965 0.00881        0.948        0.983
   11    417       2    0.961 0.00935        0.942        0.979
   12    415       2    0.956 0.00987        0.937        0.976
   13    413       1    0.954 0.01011        0.934        0.974
   14    412       3    0.947 0.01080        0.926        0.968
   15    409       2    0.942 0.01123        0.920        0.964
   16    407       2    0.938 0.01165        0.915        0.961
   17    405       3    0.931 0.01223        0.907        0.955
   18    402       3    0.924 0.01278        0.899        0.949
   19    399       2    0.919 0.01313        0.894        0.945
   20    397       5    0.907 0.01395        0.880        0.935
   21    392       2    0.903 0.01425        0.875        0.931
   22    390       1    0.900 0.01440        0.873        0.929
   23    389       1    0.898 0.01455        0.870        0.927
   24    388       4    0.889 0.01512        0.860        0.919
   25    384       3    0.882 0.01552        0.852        0.913
   26    381       3    0.875 0.01591        0.844        0.907
   27    378       2    0.870 0.01616        0.839        0.903
   28    376       2    0.866 0.01640        0.834        0.898
   30    374       2    0.861 0.01664        0.829        0.894
   31    372       1    0.859 0.01675        0.827        0.892
   32    371       2    0.854 0.01698        0.822        0.888
   33    369       2    0.850 0.01720        0.816        0.884
   34    367       2    0.845 0.01742        0.811        0.880
   35    365       4    0.836 0.01783        0.801        0.871
   36    361       3    0.829 0.01813        0.794        0.865
   37    358       4    0.819 0.01851        0.784        0.857
   38    354       1    0.817 0.01860        0.781        0.854
   39    353       2    0.812 0.01878        0.777        0.850
   40    351       4    0.803 0.01913        0.767        0.842
   42    347       2    0.799 0.01929        0.762        0.837
   43    345       4    0.789 0.01962        0.752        0.829
   44    341       2    0.785 0.01977        0.747        0.824
   45    339       2    0.780 0.01993        0.742        0.820
   46    337       4    0.771 0.02022        0.732        0.812
   47    333       1    0.769 0.02029        0.730        0.809
   48    332       2    0.764 0.02043        0.725        0.805
   49    330       5    0.752 0.02077        0.713        0.794
   50    325       3    0.745 0.02096        0.705        0.788
```

```
  52    322       4    0.736 0.02121        0.696        0.779
```

These data are in wide format. We need to transform them in to long format. Having done so, you can see that the estimates survival function is the same.

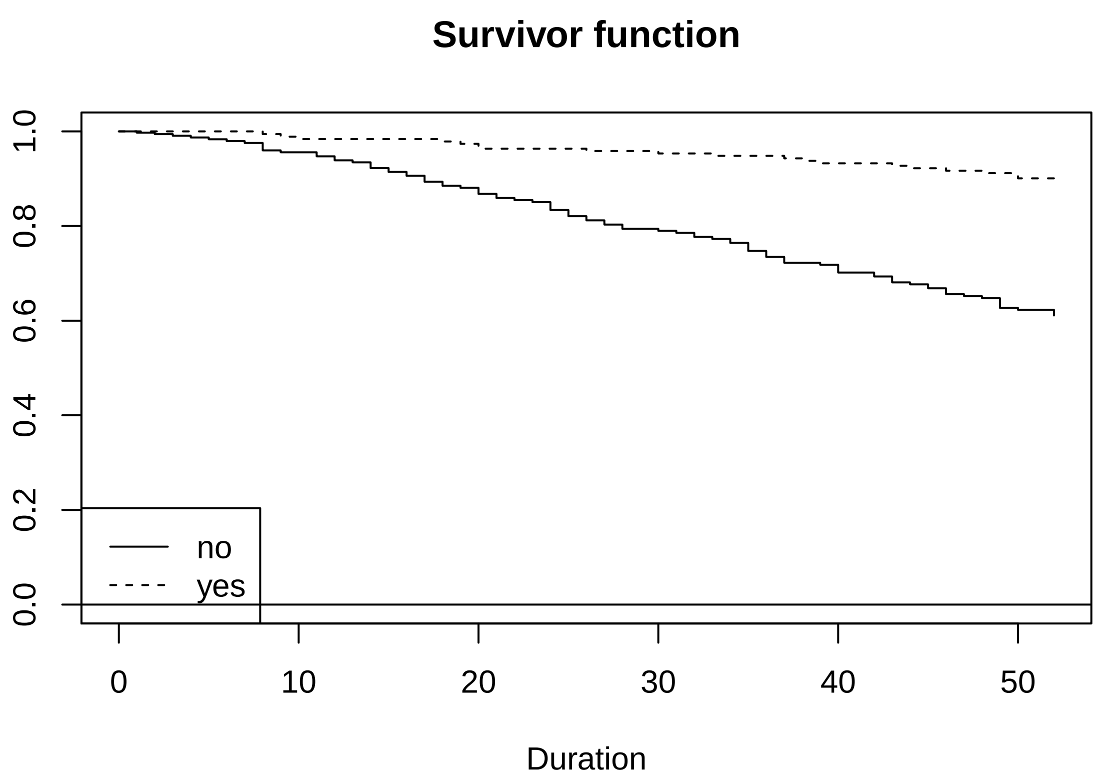
```r
# First, add an ID variable (will be useful later)
Rossi <- Rossi %>% mutate(id = row_number())

# Convert to long format
Rossi.long <- Rossi %>% gather(emp, employed, starts_with("emp")) %>% # remove missing data
filter(!is.na(employed)) %>% # calculate times at start and end of week
mutate(end = as.numeric(str_sub(emp, 4, -1)), start = end - 1) %>% # sort so easier to check visually a
# indicator
mutate(arrest_start = ifelse(arrest == 1 & week == end, 1, 0), employed = factor(employed,
    labels = c("no", "yes"))) %>% as_data_frame()

summary(survfit(Surv(Rossi.long$start, Rossi.long$end, Rossi.long$arrest_start) ~
    1, data = Rossi.long))
```

```
Call: survfit(formula = Surv(Rossi.long$start, Rossi.long$end, Rossi.long$arrest_start) ~
    1, data = Rossi.long)

 time n.risk n.event censored survival std.err lower 95% CI upper 95% CI
    1    432       1      431    0.998 0.00231        0.993        1.000
    2    431       1      430    0.995 0.00327        0.989        1.000
    3    430       1      429    0.993 0.00400        0.985        1.000
    4    429       1      428    0.991 0.00461        0.982        1.000
    5    428       1      427    0.988 0.00515        0.978        0.999
    6    427       1      426    0.986 0.00563        0.975        0.997
    7    426       1      425    0.984 0.00607        0.972        0.996
    8    425       5      420    0.972 0.00791        0.957        0.988
    9    420       2      418    0.968 0.00852        0.951        0.984
   10    418       1      417    0.965 0.00881        0.948        0.983
   11    417       2      415    0.961 0.00935        0.942        0.979
   12    415       2      413    0.956 0.00987        0.937        0.976
   13    413       1      412    0.954 0.01011        0.934        0.974
   14    412       3      409    0.947 0.01080        0.926        0.968
   15    409       2      407    0.942 0.01123        0.920        0.964
   16    407       2      405    0.938 0.01165        0.915        0.961
   17    405       3      402    0.931 0.01223        0.907        0.955
   18    402       3      399    0.924 0.01278        0.899        0.949
   19    399       2      397    0.919 0.01313        0.894        0.945
   20    397       5      392    0.907 0.01395        0.880        0.935
   21    392       2      390    0.903 0.01425        0.875        0.931
   22    390       1      389    0.900 0.01440        0.873        0.929
   23    389       1      388    0.898 0.01455        0.870        0.927
   24    388       4      384    0.889 0.01512        0.860        0.919
   25    384       3      381    0.882 0.01552        0.852        0.913
   26    381       3      378    0.875 0.01591        0.844        0.907
   27    378       2      376    0.870 0.01616        0.839        0.903
   28    376       2      374    0.866 0.01640        0.834        0.898
   30    374       2      746    0.861 0.01664        0.829        0.894
   31    372       1      371    0.859 0.01675        0.827        0.892
   32    371       2      369    0.854 0.01698        0.822        0.888
   33    369       2      367    0.850 0.01720        0.816        0.884
```

```
34    367    2    365    0.845 0.01742    0.811    0.880
35    365    4    361    0.836 0.01783    0.801    0.871
36    361    3    358    0.829 0.01813    0.794    0.865
37    358    4    354    0.819 0.01851    0.784    0.857
38    354    1    353    0.817 0.01860    0.781    0.854
39    353    2    351    0.812 0.01878    0.777    0.850
40    351    4    347    0.803 0.01913    0.767    0.842
42    347    2    692    0.799 0.01929    0.762    0.837
43    345    4    341    0.789 0.01962    0.752    0.829
44    341    2    339    0.785 0.01977    0.747    0.824
45    339    2    337    0.780 0.01993    0.742    0.820
46    337    4    333    0.771 0.02022    0.732    0.812
47    333    1    332    0.769 0.02029    0.730    0.809
48    332    2    330    0.764 0.02043    0.725    0.805
49    330    5    325    0.752 0.02077    0.713    0.794
50    325    3    322    0.745 0.02096    0.705    0.788
52    322    4    640    0.736 0.02121    0.696    0.779
```

The reason for doing this is to allow time varying covariates to be included in the analysis. In this case the explanatory variable is whether the person is in employment, the outcome variable is whether the person is arrested. Now we can do regressions using these data, but lets start with a KM plot.

```
r.surv <- Surv(Rossi.long$start, Rossi.long$end, Rossi.long$arrest_start)
plot(r.surv, strata = Rossi.long$employed, fn = "surv", conf = NULL)
```



This suggests that people in employment are less likely to be arrested. Let's try a regression.

```r
mod1 <- phreg(r.surv ~ employed, data = Rossi.long, shape = 1)
summary(mod1)
```

```
Call:
phreg(formula = r.surv ~ employed, data = Rossi.long, shape = 1)

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
employed
              no    0.532     0         1           (reference)
             yes    0.468    -1.421     0.242     0.246     0.000

log(scale)                    4.719               0.103     0.000

 Shape is fixed at  1

Events                    114
Total time at risk         19809
Max. log. likelihood      -680.36
LR test statistic         43.24
Degrees of freedom        1
Overall p-value           4.85212e-11
```

```r
mod2 <- phreg(r.surv ~ strata(employed), data = Rossi.long)
summary(mod2)
```

```
Call:
phreg(formula = r.surv ~ strata(employed), data = Rossi.long)

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
log(scale):1                  4.446               0.092     0.000
log(shape):1                  0.350               0.091     0.000
log(scale):2                  5.340               0.359     0.000
log(shape):2                  0.463               0.238     0.052

Events                    114
Total time at risk         19809
Max. log. likelihood      -672.12
```

```r
-2 * (mod1$loglik[2] - mod2$loglik[2])
```

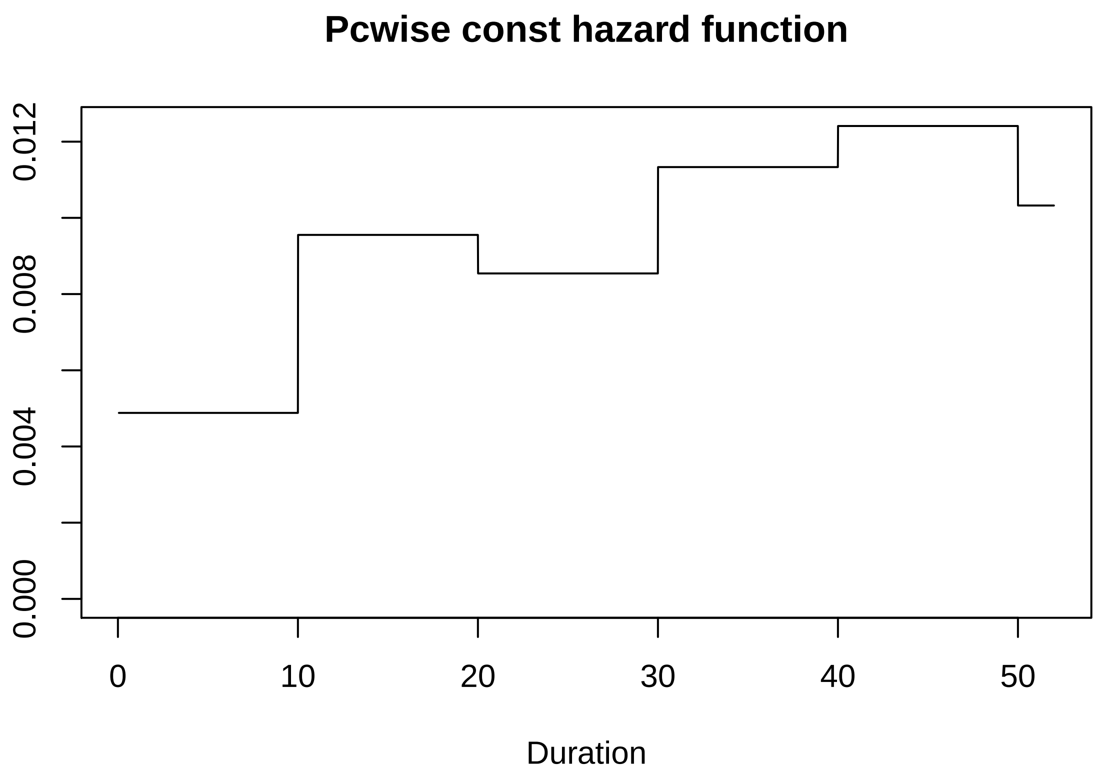
```
[1] 16.47515
```

```r
mod3 <- phreg(r.surv ~ employed, data = Rossi.long, dist = "pch", cuts = seq(10,
    50, by = 10))
summary(mod3)
```

```
Call:
phreg(formula = r.surv ~ employed, data = Rossi.long, dist = "pch",
    cuts = seq(10, 50, by = 10))

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
employed
              no    0.532     0         1           (reference)
             yes    0.468    -1.482     0.227     0.247     0.000
```

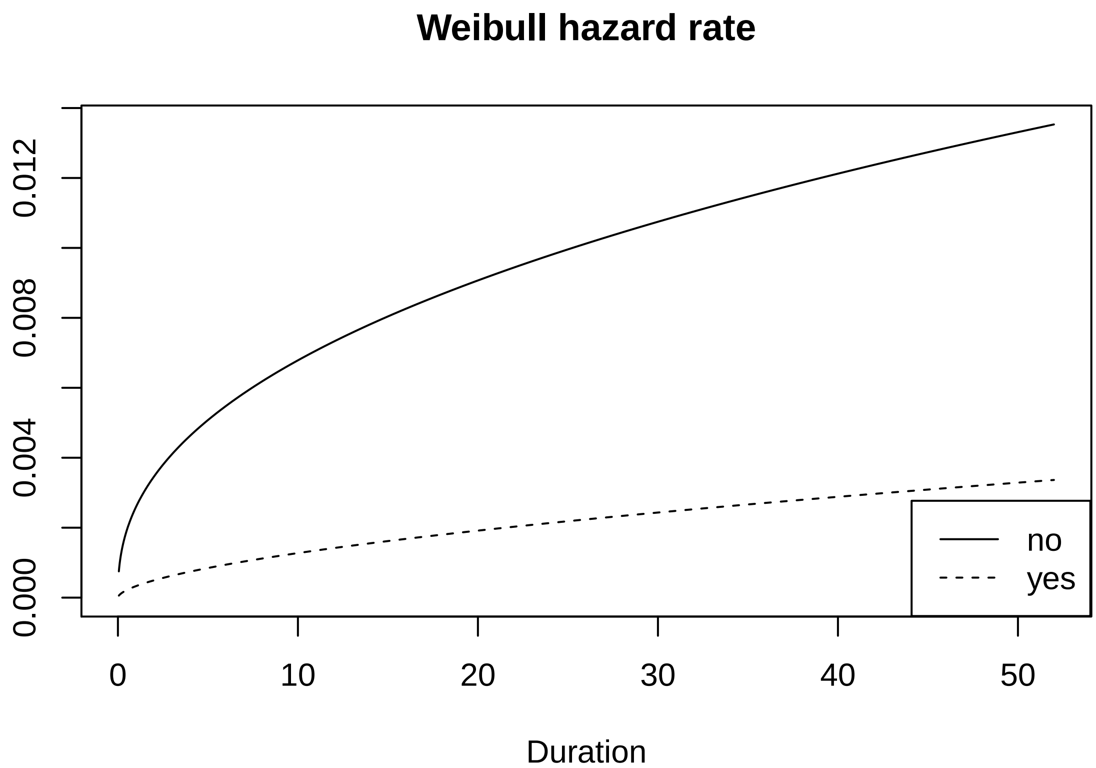
```
Events                    114
Total time at risk         19809
Max. log. likelihood      -675.02
LR test statistic         47.03
Degrees of freedom        1
Overall p-value           7.00184e-12
```

```r
plot(mod3, fn = "haz")
```



No real evidence of a non-monotonic hazard function, so Weibull looks best. Plot the two hazard functions to get a visual impression of the size of the difference.

```r
plot(mod2, fn = "haz", main = "Weibull hazard rate")
```



Homework

Focus on event count analysis; event history analysis is a level up in difficulty!

1. Use the dataset `NMES1988`, which is in the `AER` package (which you will need to install if you haven't already). Have a look at the help page for details.
2. The outcome variable of interest is `visits`, the number of visits to a doctor. Try plotting a histogram of this variable.
3. Try a poisson regression using one or more of the following variable as explanatory variables: hospital, health, chronic, gender, school, and insurance.
4. Make sure you can interpret the results. For example, how many more (or less) hospital visits are made by a typical man than a typical woman?
5. Is there evidence of overdispersion?