# Advanced Quantitative Methods: Limited Dependent Variables

Johan A. Elkink

University College Dublin

1 April 2014

# Components

Two components of the model:

$$\begin{array}{l|l} \mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2) & \text{Stochastic} \\ \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} & \text{Systematic} \end{array}$$

Generalised version (not necessarily linear):

$$\begin{array}{l|l} \mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha}) & \text{Stochastic} \\ \boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta}) & \text{Systematic} \end{array}$$

(King 1998, 8)

# Components

$$
\begin{array}{c|l}
\mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha}) & \text{Stochastic} \\
\boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta}) & \text{Systematic}
\end{array}
$$

**Stochastic** component: varies over repeated (hypothetical) observations on the same unit.

**Systematic** component: varies across units, but constant given **X**.

(King 1998, 8)

# Uncertainty

$$\begin{array}{l|l} \mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha}) & \text{Stochastic} \\ \boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta}) & \text{Systematic} \end{array}$$

Two types of uncertainty:

**Estimation uncertainty**: lack of knowledge about $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$; can be reduced by increasing *n*.
**Fundamental uncertainty**: represented by stochastic component and exists independent of researcher.

# Levels of measurement

| | Discreet | Continuous |
|---|---|---|
| **Nominal** | party choice | - |
| Ordinal | | - |
| | | - |
| Interval | | |
| Ratio | | |

Categories in no particular order

(examples in cells)

# Levels of measurement

| | Discreet | Continuous |
|---|---|---|
| Nominal | party choice | - |
| **Ordinal** | education level | - |
| | | - |
| Interval | | |
| Ratio | | |

Categories in a specific order

(examples in cells)

# Levels of measurement

| | Discreet | Continuous |
|---|---|---|
| Nominal | party choice | - |
| Ordinal | education level | - |
| | | - |
| **Interval** | how likely to vote ... | temperature |
| Ratio | | |

All values possible

(examples in cells)

# Levels of measurement

| | Discreet | Continuous |
|---|---|---|
| Nominal | party choice | - |
| Ordinal | education level | - |
| | | - |
| Interval | how likely to vote ... | temperature |
| **Ratio** | deaths in war | ideological distance |

All values possible, with a meaningful zero point

(examples in cells)

# Levels of measurement

| | Discreet | Continuous |
|---|---|---|
| Nominal | party choice | - |
| Ordinal | education level | - |
| **Binary** | turnout | - |
| Interval | how likely to vote ... | temperature |
| Ratio | deaths in war | ideological distance |

Two categories, coded as 0 and 1

(examples in cells)

# Limited dependent variables

When a dependent variable is not continuous, or is truncated for some reason, a linear model would lead to implausible predictions.

E.g. regressing whether someone voted on a set of independent variables will give somewhat reasonable estimates (see previous homework), but using these estimates to calculate predictions leads to **meaningless predictions**.

# Limited dependent variables

When a dependent variable is not continuous, or is truncated for some reason, a linear model would lead to implausible predictions.

E.g. regressing whether someone voted on a set of independent variables will give somewhat reasonable estimates (see previous homework), but using these estimates to calculate predictions leads to **meaningless predictions**.

Furthermore, estimating limited dependent variable data with a linear model implies serious **heteroscedasticity**.

# Outline

1. Binary
2. Multiple categories
3. Count
4. Survival

# Binary models

Binary models have a dependent variable consisting of two categories.

For example,

- Vote on a particular law
- Turning out in an election
- Approval in a referendum
- Bankrupt or not

# Logistic regression

Stochastic:

$$y_i = Bernoulli(\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

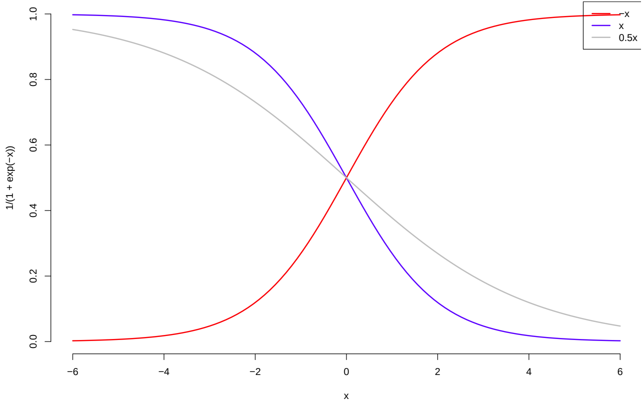
Variance of stochastic component:

$$V(y_i) = \pi_i(1-\pi_i)$$

Systematic:

$$\pi_i = g(\mathbf{x}_i\boldsymbol{\beta}) = \frac{1}{1+e^{-\mathbf{x}_i\boldsymbol{\beta}}}$$

# Logistic distribution



# Logistic regression

Resulting predictions:

- are bounded between 0 and 1
- show limited effect at extremes
- are a smooth, monotone translation from linear prediction $\mathbf{x}_i\boldsymbol{\beta}$

# Loglikelihood function

Systematic part:

$$\pi_i = \frac{1}{1 + e^{-\mathbf{x}_i \boldsymbol{\beta}}}$$

Stochastic part:

$$\begin{aligned}
P(y_i = 1) &= \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \\
f(\mathbf{y}|\pi_{\mathbf{i}}) &= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \\
\ell(\mathbf{y}) &= \log \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \\
&= \sum_{i=1}^{n} y_i \log \pi_i + \sum_{i=1}^{n}(1 - y_i) \log(1 - \pi_i)
\end{aligned}$$

# Exercise

Using the `asiabaro.dta` data set, estimate a logistic regression explaining abstention in elections by trust in the government, satisfaction with democracy, gender, urbanisation, and preference for democracy.

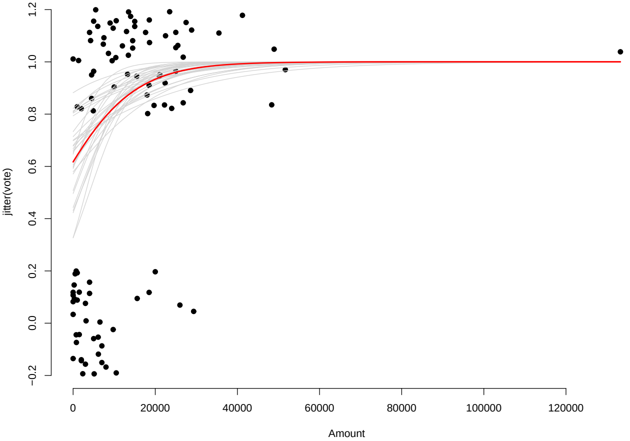
# Graphical interpretation

Because of the nonlinear relation between $\mathbf{x}_i\boldsymbol{\beta}$ and $\mathbf{y}_i$, additional tools can aid the interpretation of the size of an effect beyond just looking at $\hat{\boldsymbol{\beta}}$.

One method is to plot the relationship between one $\mathbf{x}$ and $\pi$, holding the other values of $\mathbf{X}$ constant (e.g. at the mean, median, etc.).

Because the **link function** $g(\mathbf{X}\beta)$ is not linear (but instead $g(\mathbf{X}\beta) = \frac{1}{1+e^{-\mathbf{X}\boldsymbol{\beta}}}$), the effect of $\mathbf{X}$ on $\mathbf{y}$ depends on all $\mathbf{X}$.

# Graphical: example



# Fitted values

A second useful way of interpreting **logit** regression coefficients is by describing typical cases or interesting examples.

| Party | Amount | $P(Vote = 1)$ | 95% C.I. |
|---|---|---|---|
| Republican | $10000 | .83 | [.69,.95] |
| Democrat | $10000 | .44 | [.27,.65] |
| Republican | $20000 | .93 | [.84,.99] |
| Democrat | $20000 | .69 | [.44,.95] |

# First differences

Basic idea is to calculate and present:

$$\Delta_{\hat{\pi}} = g(\mathbf{X}\beta) - g(\mathbf{X}^*\beta),$$

whereby $\mathbf{X}^*$ differs only in one variable from $\mathbf{X}$.

| Variable | Values | Diff | 95% C.I. |
|---|---|---|---|
| Party | Rep, Dem | -.34 | [-.53,-.11] |
| Amount | $10000, $20000 | .11 | [.03,.20] |

# Derivatives

$$\frac{\partial \hat{\pi}}{\partial \mathbf{x}_j} = \beta_j \hat{\pi}(1 - \hat{\pi})$$

Hence a quick method to interpret logit coefficients is to divide them by 4 to get the slope at $\hat{\pi} = 0.5$.

# Presentation

Bottom line: it is much better to present *interpretable and understandable* inferences, with an indication of the level of *uncertainty*, than to present simply estimated coefficients.

# Presentation

E.g. "An increase in automobile support for a Republican senator from \$10000 to \$20000 in total increases his or her probability to vote for the Corporate Average Fuel Economy standard bill by 11%, give or take 7%, all else equal."

# Confusion matrix

Evaluating the performance of the binary model can be done by using the **confusion matrix**:

| | | True value 1 | True value 0 | |
|---|---|---|---|---|
| Prediction | 1 | True positive (TP) | False positive (FP) | |
| Prediction | 0 | False negative (TN) | True negative (FN) | |
| | | | | |

# Confusion matrix

Evaluating the performance of the binary model can be done by using the **confusion matrix**:

| | | True value 1 | True value 0 | |
|---|---|---|---|---|
| Prediction | 1 | True positive (TP) | False positive (FP) | Precision: $\frac{TP}{TP+FP}$ |
| | 0 | False negative (TN) | True negative (FN) | |
| | | Sensitivity: $\frac{TP}{TP+FN}$ | Specificity: $\frac{TN}{FP+TN}$ | |

# Receiver Operating Characteristic curve

The accuracy of predictions will depend on the threshold probability – variations on default of $\hat{\pi} = 0.5$ are possible.

Depending on the application, it might be better or worse to over- or underestimate ones relative to zeros.

The ROC-curve plots, for all possible thresholds, the true positive rate against the false positive rate.

An ROC-curve closer to the 45 degree line indicates a better predictive performance; any predictions under this line indicate worse than random prediction.

# Area Under Curve (AUC)

Given the above, we can also calculate the area under the ROC-curve as a measure of prediction quality. This is somewhat related to the Gini coefficient for income distributions ($G = 2AUC - 1$).

An extensive example in R can be found at `http://www.joselkink.net/pub/ROC-example.html`.

# Exercise

Using the logistic previous logistic regression:

1. What is the slope of the regression line with respect to preference for democracy at $\hat{\pi} = .5$?
2. Calculate predicted differences for urban vs rural respondent.
3. Calculate first differences for urban vs rural and female vs male.
4. Plot predicted probabilities as a function of preference for democracy.
5. Calculate the AUC score for this model.

# Threshold models

Imagine we have a latent, unobserved variable:

$$\mathbf{y}^* \sim f(\boldsymbol{\mu})$$
$$\boldsymbol{\mu} = \mathbf{X}\beta$$

With the following observation mechanism:

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0. \end{cases}$$

# Threshold models

If $f(\mu_i)$ is the **standardised logistic distribution**, we replicate the logit model.

$$f(\mathbf{x}_i\beta) = \frac{e^{y_i^* - \mu_i}}{(1 + e^{y_i^* - \mu_i})^2}$$

# Threshold models

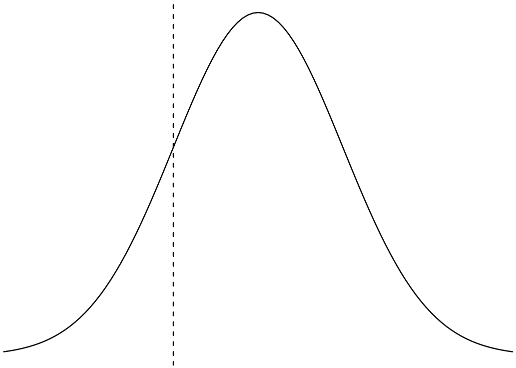
Alternatively, if $f(\mu_i)$ is the **cumulative standard normal distribution** ($\sigma^2 = 1$), we have the **probit** model.

Hence $\boldsymbol{\beta}$ can now be interpreted as the effect of an increase by 1 in $\mathbf{x}$ on $\mathbf{y}^*$, whereby the unit of $\mathbf{y}^*$ is one standard deviation.

$$P(y_i = 1|\boldsymbol{\beta}, \mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\beta}),$$

whereby $\Phi(x)$ is the cumulative standard normal distribution (i.e. the surface between $-\infty$ and $x$ under a normal distribution with $\mu = 0$ and $\sigma^2 = 1$).

# Threshold models



# Loglikelihood function

Systematic part:

$$\pi_i = \Phi(\mathbf{x}_i\boldsymbol{\beta}),$$

where $\Phi(x)$ is the **cumulative standard normal distribution**.

Stochastic part:

$$P(y_i = 1) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$f(\mathbf{y}|\pi_{\mathbf{i}}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\ell(\mathbf{y}) = \sum_{i=1}^{n} y_i \log \pi_i + \sum_{i=1}^{n}(1-y_i)\log(1-\pi_i)$$

# Exercise

Repeat the previous estimation using probit instead of logit.

1. Plot predicted probabilities with respect to preference for democracy.
2. Evaluate the predictive performance of this model.
3. Add `suitdemoc` as a dependent variable to the model. Does the predictive performance improve?

# Outline

# Ordered data

Models where the dependent variable is categorical, and the categories are in a particular order.

We can therefore take a similar latent variable approach.

# Ordered probit

$$y_i^* \sim N(\mu_i, 1)$$
$$\mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

With the following observation mechanism:

$$y_i = j \quad \text{if} \quad \tau_{j-1} \leq y_i^* \leq \tau_j$$

# Ordered probit

$$y_i^* \sim N(\mu_i, 1)$$
$$\mu_i = \mathbf{x}_i \boldsymbol{\beta}$$

With the following observation mechanism:

$$y_i = j \quad \text{if} \quad \tau_{j-1} \leq y_i^* \leq \tau_j$$

Or, alternatively, when we use dummy variables for each category:

$$y_{ij} = \begin{cases} 1 & \text{if} \quad \tau_{j-1} \leq y_i^* \leq \tau_j \\ 0 & \text{otherwise.} \end{cases}$$

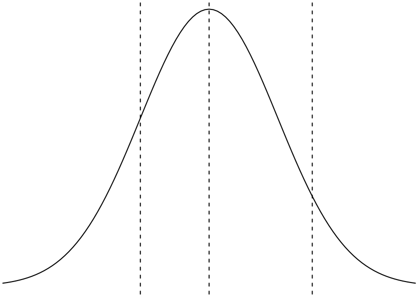
Note that there are no lower or upper bounds for the first and last category, respectively.

# Ordered probit

$$y_{ij} = \begin{cases} 1 & \text{if} \quad \tau_{j-1} \leq y_i^* \leq \tau_j \\ 0 & \text{otherwise.} \end{cases}$$

$$P(y_i = j|\beta, \mathbf{x}_i) = \Phi(\tau_j - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\tau_{j-1} - \mathbf{x}_i\boldsymbol{\beta})$$

# Ordered probit



# Ordered probit

To estimate in R:

```
library(MASS)
m <- polr(y ~ x1 + x2, method="probit")
```

# Ordered probit

To estimate in R:

```
library(MASS)
m <- polr(y ~ x1 + x2, method="probit")
```

To get predicted probabilities:

```
pnorm(m$zeta[2] - xb) - pnorm(m$zeta[1] - xb)
predict(m, type="probs")
```

# Loglikelihood function

Assuming $y_i$ can take the values $0, 1, 2, ..., p$.

$$P(y_i = 0) = \Phi(\tau_1 - \mathbf{x}_i\boldsymbol{\beta})$$
$$P(y_i = p) = \Phi(\mathbf{x}_i\boldsymbol{\beta} - \tau_p)$$
$$P(y_i = j|0 < y_i < p) = \Phi(\tau_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\tau_j - \mathbf{x}_i\boldsymbol{\beta})$$

$$\ell(\mathbf{y}) = \sum_{y_i=0} \log(\Phi(\tau_1 - \mathbf{x}_i\boldsymbol{\beta})) + \sum_{y_i=p} \log(\Phi(\mathbf{x}_i\boldsymbol{\beta} - \tau_p)) + \sum_{0<y_i<p} \log(\Phi(\tau_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\tau_j - \mathbf{x}_i\boldsymbol{\beta}))$$

# Exercise

Using the `asiabaro.dta` data set, estimate a model explaining interest in politics by gender, education, and reliance on fate.

1. Perform *t*-tests for each of the independent variables.
   Use: `se <- sqrt(diag(vcov(m)))`
2. Check the standard errors of the $\tau$ values – are the categories clearly separated?
3. Calculate the predicted probability for males and females of the respondent being very interested in politics.

# Multinomial model

Here the dependent variable consists of multiple categories, without particular order.

A latent variable approach will therefore not work.

The probabilities for a particular case to be in any of those categories still has to be one.

# Multinomial logit

R function: *multinom*, in package *nnet*.

```
m <- multinom(y ~ x1 + x2)
summary(m)
```

```
predict(m, type="probs")
```

# Multinomial logit

$$P(y_i = j|\beta, \mathbf{x}_i) = \begin{cases} \frac{1}{\sum_{k=1}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_k}} & \text{if} \quad j = 1 \\ \frac{e^{\mathbf{x}_i \boldsymbol{\beta}_j}}{\sum_{k=1}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_k}} & \text{if} \quad j > 1 \end{cases}$$

# Multinomial logit

$$P(y_i = j|\beta, \mathbf{x}_i) = \begin{cases} \frac{1}{\sum_{k=1}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_k}} & \text{if} \quad j = 1 \\ \frac{e^{\mathbf{x}_i \boldsymbol{\beta}_j}}{\sum_{k=1}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_k}} & \text{if} \quad j > 1 \end{cases}$$

To get predicted probabilities for a new dataset ($\mathbf{X}^*$) - e.g. one that is constant on all variables but one:

```
m <- multinom(y ~ x1 + x2 + x3 + x4)
Xb <- X %*% t(coef(m))
denominator <- 1 - rowSums(exp(Xb))
probs <- exp(Xb) / denominator
baseline <- 1 - rowSums(p)
p <- cbind(baseline, p)
```

# Loglikelihood function

$$P(y_i = j|\beta, \mathbf{x}_i) = \begin{cases} \frac{1}{\sum_{k=1}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_k}} & \text{if} \quad j = 1 \\ \frac{e^{\mathbf{x}_i \boldsymbol{\beta}_j}}{\sum_{k=1}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_k}} & \text{if} \quad j > 1 \end{cases}$$

$$\ell(\mathbf{y}) = \sum_{i=1}^{n} \sum_{j=0}^{p} \left[ I(y_i = j)\mathbf{x}_i \boldsymbol{\beta}_j - \log \left( \sum_{j=0}^{p} e^{\mathbf{x}_i \boldsymbol{\beta}_j} \right) \right]$$

# Exercise

Using data set `asiabaro.dta`, selecting only data from Taiwan, explain party choice by urbanisation, gender, age, and trust in the police. Estimate a multinomial model and interpret the results.

# Outline

# Count models

Here the dependent variable is a count (of events).

For example,

- The number of conflicts in a particular period
- The number of coups d'état in the 1980s
- The number of visits to a psychologist for a respondent

# Count models

Here the dependent variable is a count (of events).

For example,

- The number of conflicts in a particular period
- The number of coups d'état in the 1980s
- The number of visits to a psychologist for a respondent

Note:

- Truncated at zero (no negative outcome possible)

# Count models

Here the dependent variable is a count (of events).

For example,

- The number of conflicts in a particular period
- The number of coups d'état in the 1980s
- The number of visits to a psychologist for a respondent

Note:

- Truncated at zero (no negative outcome possible)
- No upper limit

# Count models

Here the dependent variable is a count (of events).

For example,

- The number of conflicts in a particular period
- The number of coups d'état in the 1980s
- The number of visits to a psychologist for a respondent

Note:

- Truncated at zero (no negative outcome possible)
- No upper limit
- Limited by time, place, or both

# Poisson regression

Stochastic: $y_i \sim Poisson(\lambda_i)$

Systematic: $\lambda_i = e^{\mathbf{x}_i \boldsymbol{\beta}}$

Poisson distribution:

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

Note that there is no parameter for the variance. This leads to regular underestimation (**overdispersion**, most common) or overestimation of the variance (**underdispersion**).

```
glm(y ~ x1 + x2 + x3, family=poisson)
```

# Poisson regression

The exponential of the coefficients can be interpreted in a multiplicative sense.

E.g. if the coefficient of $\mathbf{x}_1$ is 0.012, then $e^{0.012} = 1.012$ implies that an increase of $\mathbf{x}_1$ by 1 increases $\mathbf{y}$ by 1.2%.

# Estimating overdispersion

The estimated overdispersion of a Poisson model is

$$\frac{1}{n-k}\sum_{i}^{n}\left(\frac{y_i-\hat{y}_i}{sd(\hat{y}_i)}\right)^2 = \frac{1}{n-k}\sum_{i}^{n}\left(\frac{y_i-e^{X_i\beta}}{\sqrt{e^{X_i\beta}}}\right)^2,$$

which has a $\chi^2_{n-k}$ distribution.

# Negative binomial

The negative binomial model is useful for **overdispersed** data:

Stochastic: $y_i \sim NegBin(\phi, \sigma^2)$

Systematic: $\phi = e^{\mathbf{x}_i \boldsymbol{\beta}} = E(y_i)$

Variance: $V(\mathbf{y}|\mathbf{X}) = \phi\sigma^2$

When $\sigma^2$ approaches 1, the negative binomial approaches the Poisson distribution.

```
library(MASS)
glm.nb(y ~ x1 + x2 + x3)
```

# Exercise

Open dataset *quine* which is part of the MASS library in R.

| | |
|---|---|
| **Eth** | Ethnicity: aboriginal or not |
| **Sex** | Sex of respondent |
| **Age** | Age group |
| **Lrn** | Average or slow learner |
| **Days** | Days absent from school in a year |

Estimate / interpret / check both poisson and negative binomial:
$Days \sim Eth + Sex + Age + Lrn$

# Outline

# Concept

Basic idea is to estimate the duration of something, or the time until death or failure.

# Survivor function

Say, our dependent variable, $\mathbf{y}$, records length of life, thus a random variable between 0 and $\infty$.

The cumulative distribution function of $\mathbf{y}$ is $F(t) = P(\mathbf{y} < t)$, i.e. the probability of death before time (younger than) $t$.

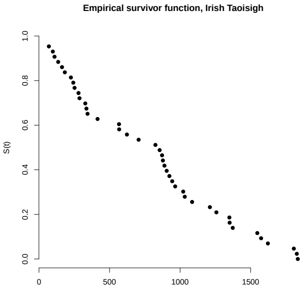
More commonly used is its complement, the probability of death later (older) than $t$: $S(t) = P(\mathbf{y} > t) = 1 - P(\mathbf{y} < t) = 1 - F(t)$. The latter is known as the **survivor function**.

Note that $S(0) = 1$ and $S(\infty) = 0$, and $S(t)$ decreases monotonically between 0 and $\infty$.

# Empirical survivor function

$$S(t) = \frac{\text{number of observations} > t}{n} = \frac{1}{n}\sum_{i}^{n} I_{(t,\infty)}(y_i),$$

whereby $I_{(a,b)}(x)$ is an indicator function which is 1 if $x$ is between $a$ and $b$, 0 otherwise.



# Hazard function

We have $S(t) = P(\mathbf{y} > t) = 1 - F(t)$. What we are usually interested in is the **hazard function**, the probability of death "now", $P(t < \mathbf{y} < t + \Delta t)$, *given* survival up until now: $P(t < \mathbf{y} < t + \Delta t | \mathbf{y} > t)$.

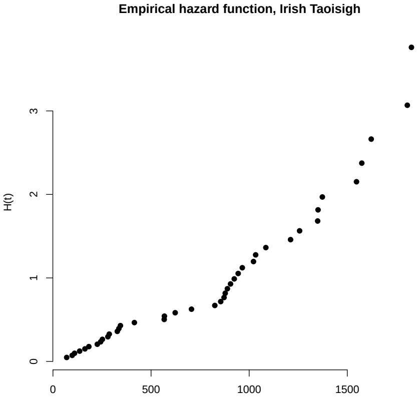
As $\Delta t$ goes to 0, this is given by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < \mathbf{y} < t + \Delta t)}{P(\mathbf{y} > t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

Alternative names: force of decrement, force of mortality, age-specific death (failure) rate, intensity function, hazard function.

# Empirical hazard function

$$H(t) = -\log_e S(t)$$



# Constant hazard model

The probability of survival is independent of the time of survival thus far.

$h(t) = \beta$

$S(t) = e^{-t\beta}$

The result is an **exponential probability model**.

# Weibull hazard model

Another typical model is the Weibull specification:

$$h(t) = \gamma\beta t^{\beta-1}$$

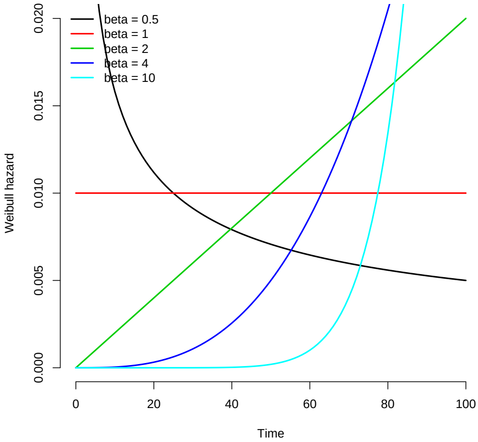
If $\beta = 1$, this reduces to:

$$h(t) = \gamma,$$

which is the exponential model.

If $\beta > 1$, the hazard increases monotonically over time; if $\beta < 1$, the hazard decreases monotonically over time.

# Weibull hazard model



# Censoring

Right censoring:

- Cases die for other reasons
- Cases die outside observed timeframe

Left censoring:

- Birth time is unknown

Note that if censoring is not independent of $Y$, estimates can be seriously biased.

# Censoring in R

A dependent variable for survival analysis is defined in R by the function *Surv*.

# Censoring in R

A dependent variable for survival analysis is defined in R by the function *Surv*.

Template 1:

```
Surv(time, event)
```

Where *time* refers to the length of time until death and *event* is 0 when right-censored, 1 when not.

# Censoring in R

A dependent variable for survival analysis is defined in R by the function *Surv*.

Template 1:

```
Surv(time, event)
```

Where *time* refers to the length of time until death and *event* is 0 when right-censored, 1 when not.

Template 2:

```
Surv(time, time2, event, type)
```

Where time is the interval [*time*, *time2*], *type* is "interval" and *event* is 0 for right-censored, 1 for normal event, 2 for left-censored, and 3 for both left- and right-censored.

# Proportional hazard model

Generally, we want to estimate survival models with independent, explanatory variables. The typical structure for this is:

$$h_x(t) = h_0(t)g(x) = h_0(t)e^{\mathbf{X}\boldsymbol{\beta}}$$

$h_0(t)$ is called the **base hazard**.

Note that the relative hazard of two cases is independent of the base hazard:

$$\frac{h_{x_1}(t)}{h_{x_2}(t)} = \frac{h_0(t)g(x_1)}{h_0(t)g(x_2)} = \frac{g(x_1)}{g(x_2)}$$

# Proportional hazard model

The base hazard can now have various different distributions, e.g. exponential.

In R:

```
summary(survreg(Surv(y,c) ~ x1 + x2 + x3,
        data=data, dist="exponential"))

summary(survreg(Surv(y,c) ~ x1 + x2 + x3,
        data=data, dist="weibull"))
```

Note that the coefficients enter multiplicatively, similar to count models. If $\beta_{\mathbf{x}_1} = -0.16$, then the multiplicative effect is $e^{-0.16} = .85$, which an increase of $\mathbf{x}_1$ by 1 leads to a 15% decrease in the hazard.

# Cox proportional hazard model

The most commonly used proportional hazard model is **nonparametric**, i.e. there is no assumption made about the distribution of $h_0$.

Using a nonparametric model leads to a slightly less efficient estimation, but a more generic one.

# Time-varying independent variables

Two types of independent variables in survival analysis can be distinguished:

- Constant over time $h_X(t) = h_0(t)e^{\mathbf{X}\boldsymbol{\beta}}$
- Varying over time $h_X(t) = h_0(t)e^{\mathbf{X}(t)\boldsymbol{\beta}}$

# Two data formats

Depending on whether there are time varying independent variables, a survival data set can be in two different formats.

Format 1:

| time | censored | $x_1$ | $x_2$ |
|---|---|---|---|
| 10 | 1 | 3 | 1 |
| 14 | 0 | 2 | 0 |
| 13 | 1 | 5 | 1 |
| 2 | 1 | 4 | 1 |

# Two data formats

Depending on whether there are time varying independent variables, a survival data set can be in two different formats.

Format 2:

| start | end | event | censored | $x_1$ | $x_2$ |
|---|---|---|---|---|---|
| 1 | 2 | 0 | 1 | 3 | 1 |
| 2 | 3 | 0 | 0 | 2 | 0 |
| 3 | 4 | 1 | 1 | 5 | 1 |
| 1 | 2 | 0 | 1 | 3 | 1 |

# Example: Recidivism

See for an extensive example, including with time varying variables:

`http://cran.r-project.org/doc/contrib/Fox-Companion/cox-regression.txt`

# Competing risks

The competing risks model is a survival model where there are multiple ways of failure or death, e.g. different causes of death.

In a competing risks model, right censoring can be included simply as another type of risk.

# Frailty

In some scenarios, one wants to assume that the baseline hazard, $h_0(t)$, varies per individual, or per group of individuals.

E.g. different types of companies might have different risks of bankruptcy.

A frailty is an extra parameter to a proportional hazard model that estimates this unit- or group-specific baseline hazard.

$$h_X(t) = \alpha h_o(t) e^{\mathbf{X}\boldsymbol{\beta}} = h_o(t) e^{X\mathbf{X}\boldsymbol{\beta} + log(\alpha)},$$

with typically $log(\alpha_i) \sim N(0, \sigma^2)$.

# Frailty in R

```
summary(survreg(Surv(y,c) ~ x1 + x2 + x3 + frailty(id,
                                            distribution="gauss")),
        data=data, dist="exponential"))
```

Whereby distribution can be "gauss" for a normally distributed frailty term, "gamma" for a gamma distribution, etc.

# Exercise: Taoisigh

Open the `taoisigh.dta` data file.

1. Estimate a baseline survival model, using an exponential hazard function, for the duration of cabinets.
2. Extend the model to see whether coalitions have a lower survival probability.
3. Estimate the model using a proportional hazard function.
4. Estimate the model with a frailty term for party.
5. What do you conclude about the impact of coalitions on the survival chances of Irish governments?