Logistic Regression

Example

Dose	No. survived	No. dead
0.0	18	7
0.5	19	6
1.0	12	13
1.5	5	20
2.0	6	19
2.5	2	23
3.0	1	24

Binary vs. continuous outcomes

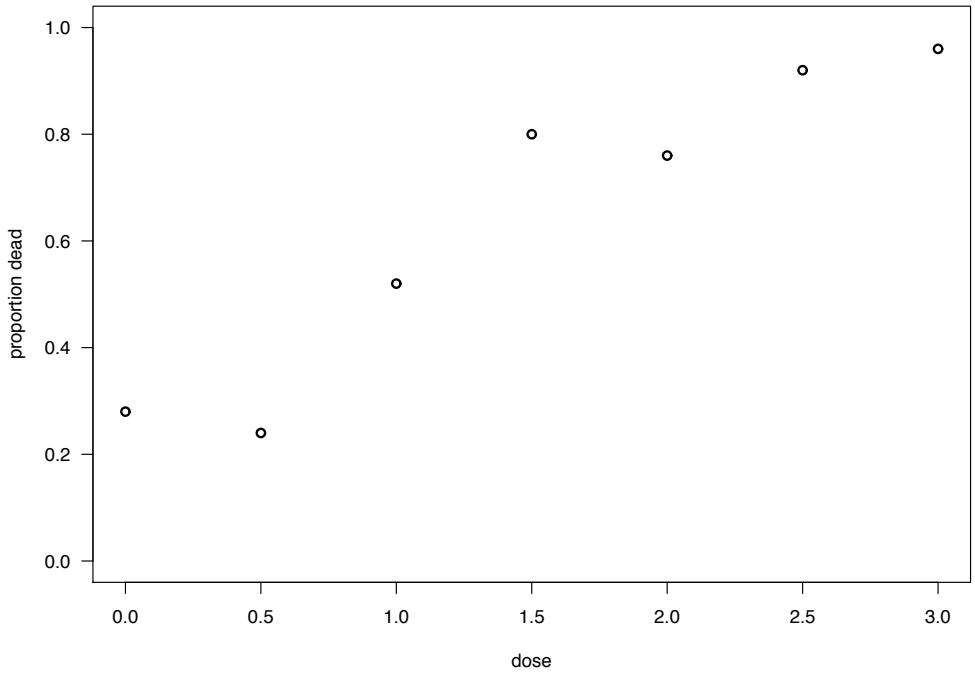
Continuous:	ANOVA	$\longleftrightarrow$	Regression
Binary:	k $\times$ 2 table	$\longleftrightarrow$	?

Goals:

$\longrightarrow$ Determine the relationship between dose and Pr(dead).

$\longrightarrow$ Find the dose at which Pr(dead) = 1/2.

A plot of the data



Linear regression

Model:

$$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim \text{ iid Normal}(0, \sigma^2)$$

This implies:

$$E(y \mid x_1, \ldots, x_k)=\beta_0 + \beta_1 x_k + \cdots + \beta_k x_k$$

$\longrightarrow$ What is the interpretation of β_i ?

Binary outcomes

Let p_d = Pr(dead | dose d)

$$p_d=\beta_0 + \beta_1 d \ ?$$

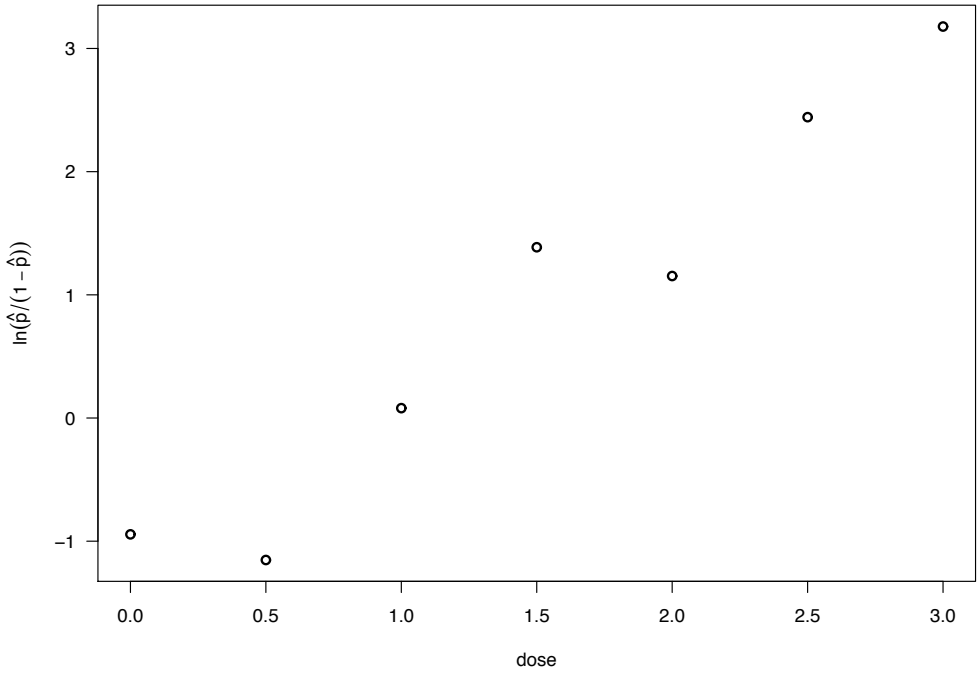
$$0 \leq p_d \leq 1 \qquad \text{but} \qquad -\infty \leq \beta_0 + \beta_1 d \leq \infty$$

Odds of death: $\quad 0 \leq \dfrac{p_d}{1-p_d} \leq \infty$

Log odds of death: $\quad -\infty \leq \ln\left(\dfrac{p_d}{1-p_d}\right) \leq \infty$

$\longrightarrow$ $\ln\left(\dfrac{p}{1-p}\right)$ is also called logit(p) or the logistic function.

logit($\hat{p}_d$) vs d



Logistic regression

$$\ln\left(\frac{p_d}{1-p_d}\right) = \beta_0 + \beta_1 d$$

Try least squares, regressing $\ln\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right)$ on the dose d?

Problems:

$\longrightarrow$ What if $\hat{p}_d = 0$ or 1?

$\longrightarrow$ SD($\hat{p}_d$) is not constant with d.

Maximum likelihood

Assume that

- $y_d \sim \text{Binomial}(n_d, p_d)$,
- y_d independent,
- $\text{logit}(p_d) = \ln(\frac{p_d}{1-p_d}) = \beta_0 + \beta_1 d$

Note: $p_d = \frac{e^{\beta_0+\beta_1 d}}{1 + e^{\beta_0+\beta_1 d}}$

Likelihood:

$$L(\beta_0, \beta_1 \mid y) = \prod_d p_d^{y_d} (1 - p_d)^{(n_d - y_d)}$$

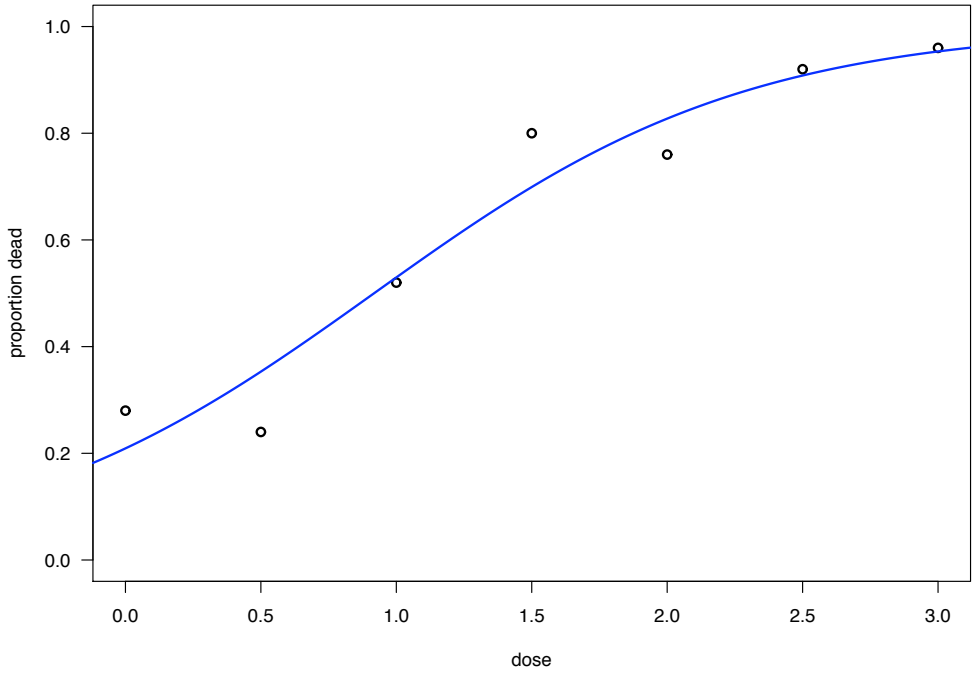
Logistic regression

Logistic regression is a special case of a *generalized linear model*.

Software output:

```
> summary(glm.out)$coef

                Est    SE  t-val   P-val
(Intercept)   -1.33  0.33  -4.06  <0.001
dose           1.44  0.23   6.29  <0.001
```

Fitted curve



Interpretation of β's

$$\ln\left(\frac{p_d}{1-p_d}\right) = \beta_0 + \beta_1 d$$

β_0 = log odds when dose = 0

Note: $\beta_0 = 0 \quad \longrightarrow \quad p_0 = \frac{1}{2}$

β_1 = change in log odds with unit increase in dose

Note: $\beta_1 = 0 \quad \longrightarrow$ survival unrelated to dose.

LD50

LD50 = dose at which Pr(dead | dose) = $\frac{1}{2}$.

$$\ln\left(\frac{1/2}{1-1/2}\right) = \beta_0 + \beta_1(\text{LD50})$$

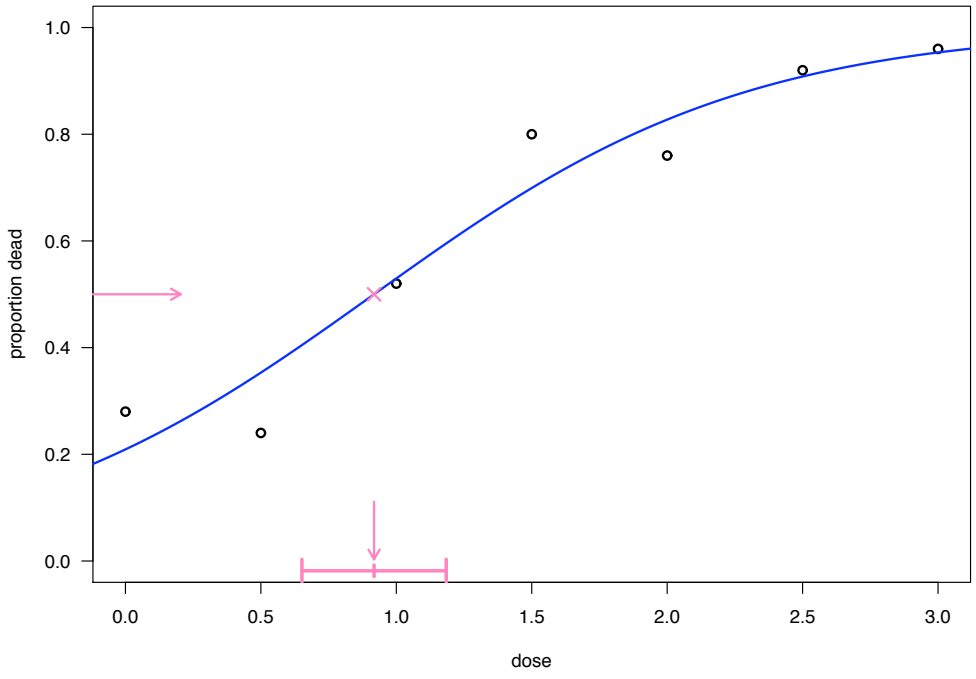
$$0 = \beta_0 + \beta_1(\text{LD50})$$

$$\text{LD50} = -\beta_0/\beta_1$$

$\widehat{\text{LD50}} = -\hat{\beta}_0/\hat{\beta}_1$

$$\hat{\text{SE}}(\widehat{\text{LD50}}) \approx |\widehat{\text{LD50}}| \sqrt{\left(\frac{\hat{\text{SE}}(\hat{\beta}_0)}{\hat{\beta}_0}\right)^2 + \left(\frac{\hat{\text{SE}}(\hat{\beta}_1)}{\hat{\beta}_1}\right)^2 - 2\,\frac{\text{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\hat{\beta}_0\,\hat{\beta}_1}}$$

LD50



Another example

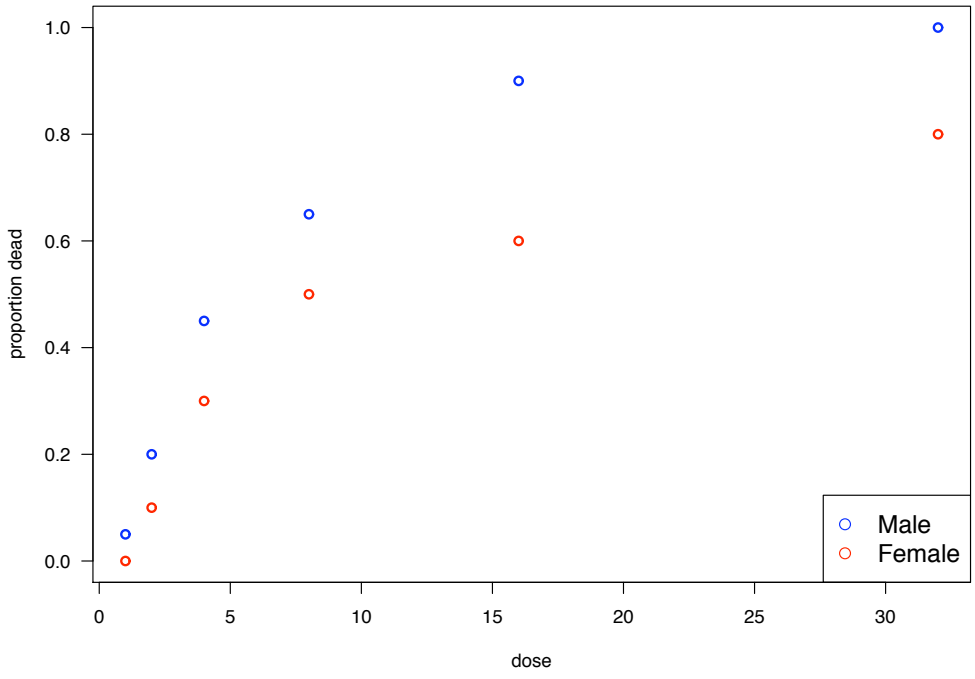
Tobacco budworm, *Heliothis virescens*

Batches of 20 male and 20 female worms were given a 3-day dose of pyrethroid *trans*-cypermethrin

The no. dead or “knocked down” in each batch was noted.

	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

A plot of the data



Analysis

Assume no sex difference

logit(p) = β_0 + $\beta_1 \times$dose

```
> summary(glm.out)$coef
                  Est     SE  t-val   P-val
  (Intercept)   -1.57   0.23   -6.8  <0.001
  dose           0.153  0.022   6.8  <0.001
```

Assume sexes completely different

logit(p) = β_0 + $\beta_1 \times$sex + $\beta_2 \times$dose + $\beta_3 \times$sex:dose

```
> summary(glm.out)$coef
                  Est     SE  t-val   P-val
  (Intercept)   -1.72   0.32   -5.3  <0.001
  sexmale       -0.21   0.51   -0.4   0.68
  dose           0.116  0.024   4.9  <0.001
  sexmale:dose   0.182  0.067   2.7   0.007
```

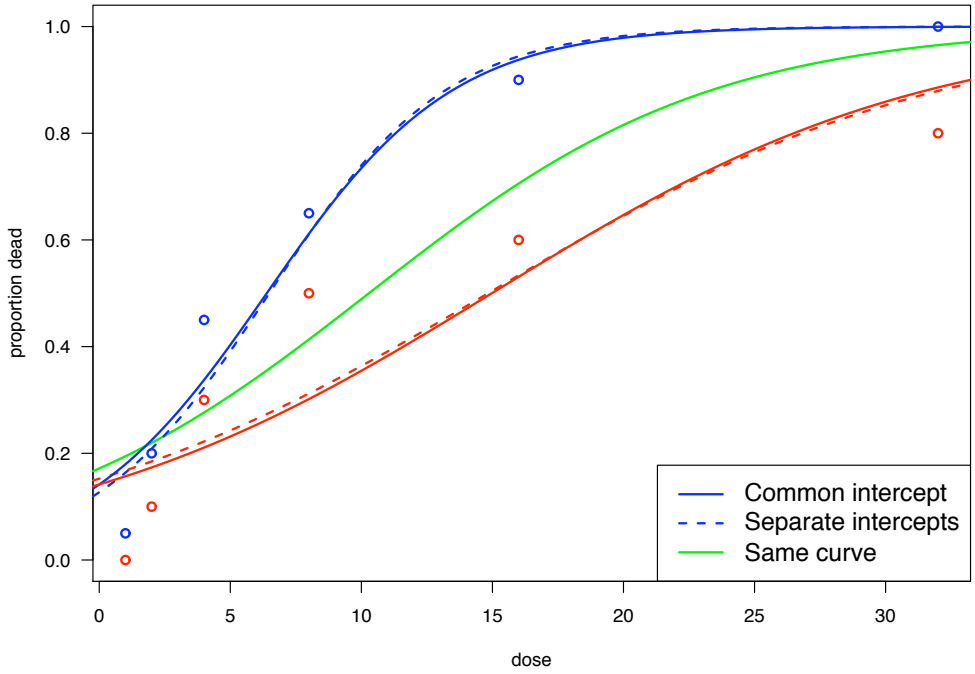
Analysis (continued)

Different slopes but common “intercept”

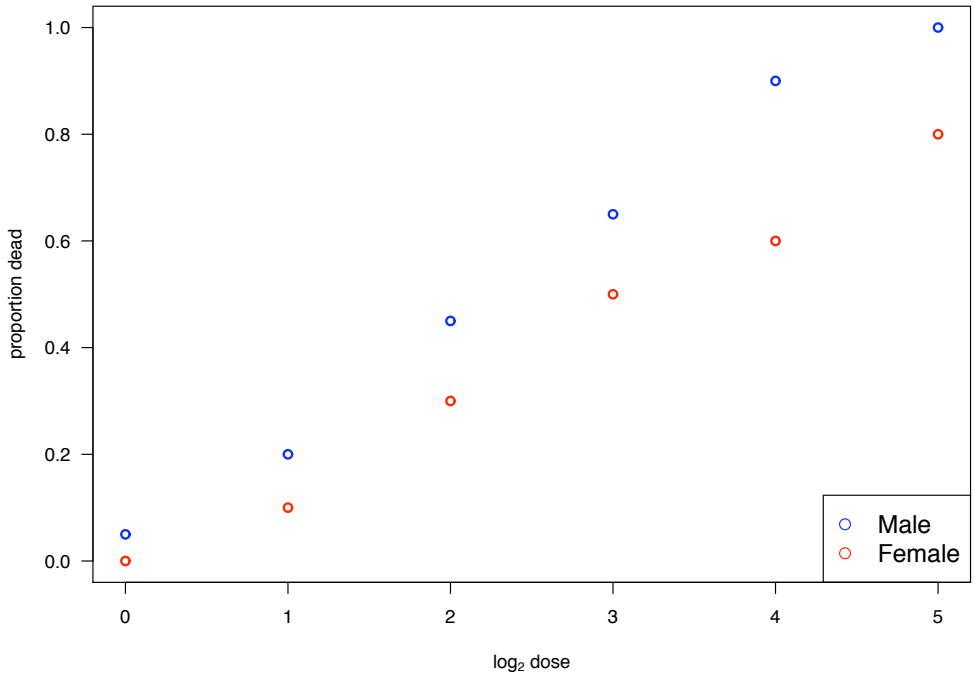
logit(p) = β_0 + $\beta_1 \times$dose + $\beta_2 \times$sex:dose

```
> summary(glm.out)$coef
                  Est     SE  t-val   P-val
  (Intercept)   -1.80   0.25   -7.2  <0.001
  dose           0.120  0.021   5.6  <0.001
  dose:sexmale   0.161  0.044   3.7  <0.001
```

Fitted curves



Plot using $\log_2$ dose



Use $\log_2$ of the dose

Assume no sex difference

logit(p) = β_0 + $\beta_1 \times \log_2$(dose)

```
> summary(glm.out)$coef
                 Est    SE  t-val   P-val
(Intercept)   -2.77  0.37   -7.6  <0.001
log2dose       1.01  0.12    8.1  <0.001
```

Assume sexes completely different

logit(p) = β_0 + $\beta_1 \times$sex + $\beta_2 \times \log_2$(dose) + $\beta_3 \times$sex:$\log_2$(dose)

```
> summary(glm.out)$coef
                     Est    SE  t-val   P-val
(Intercept)        -2.99  0.55   -5.4  <0.001
sexmale             0.17  0.78   -0.2   0.82
log2dose            0.91  0.17    5.4  <0.001
sexmale:log2dose    0.35  0.27    1.3   0.19
```

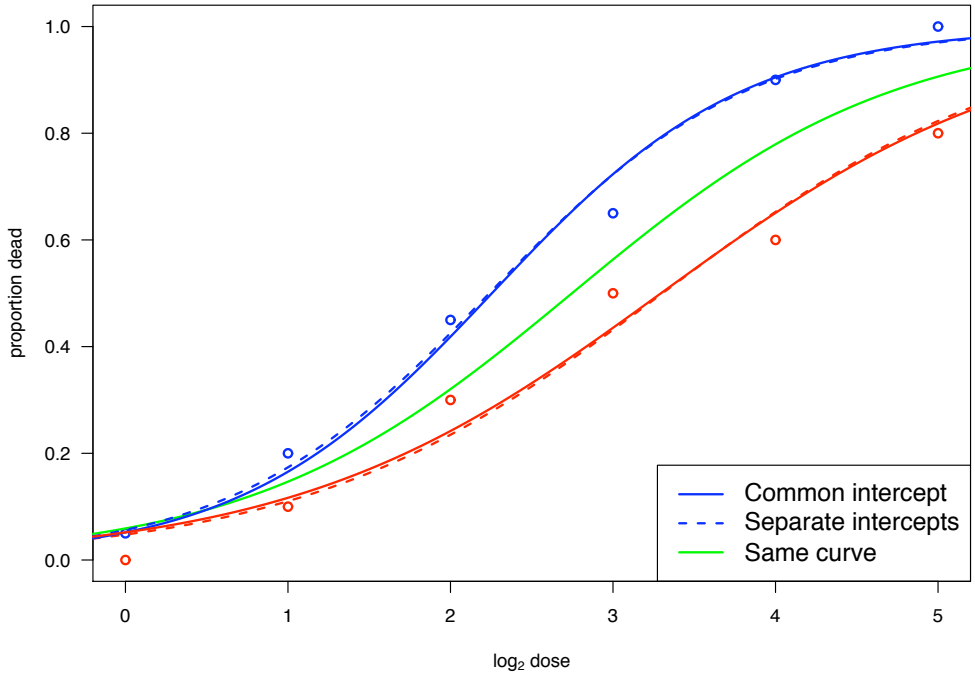
Use $\log_2$ of the dose (continued)

Different slopes but common “intercept”

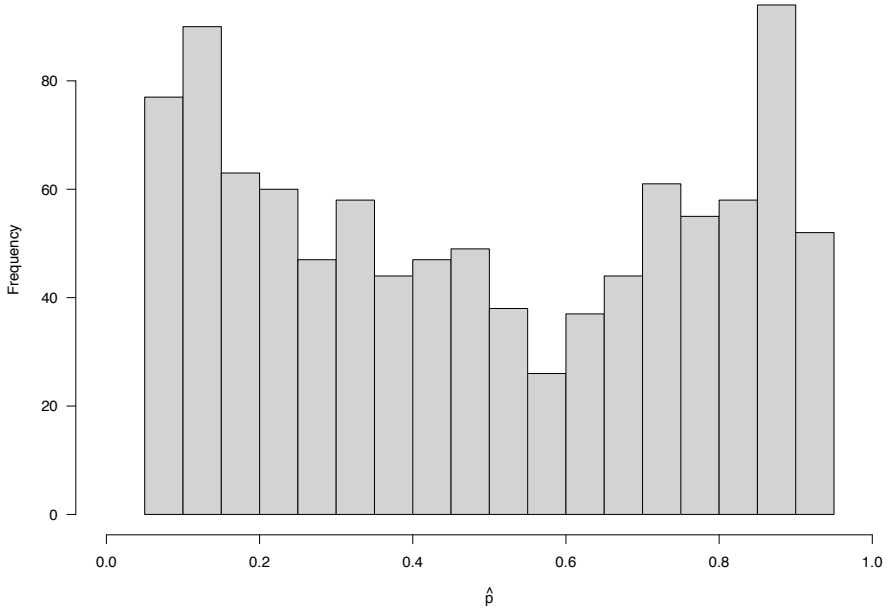
logit(p) = β_0 + $\beta_1 \times \log_2$(dose) + $\beta_2 \times$sex:$\log_2$(dose)

```
> summary(glm.out)$coef
                     Est    SE  t-val   P-val
(Intercept)        -2.91  0.39   -7.5  <0.001
log2dose            0.88  0.13    6.9  <0.001
log2dose:sexmale    0.41  0.12    3.3   0.001
```

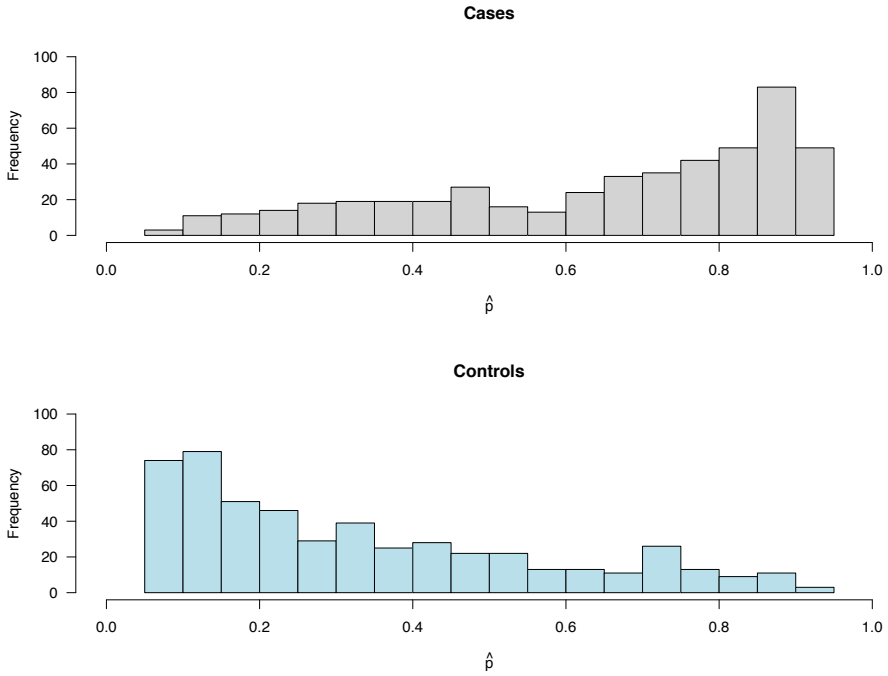
Fitted curves



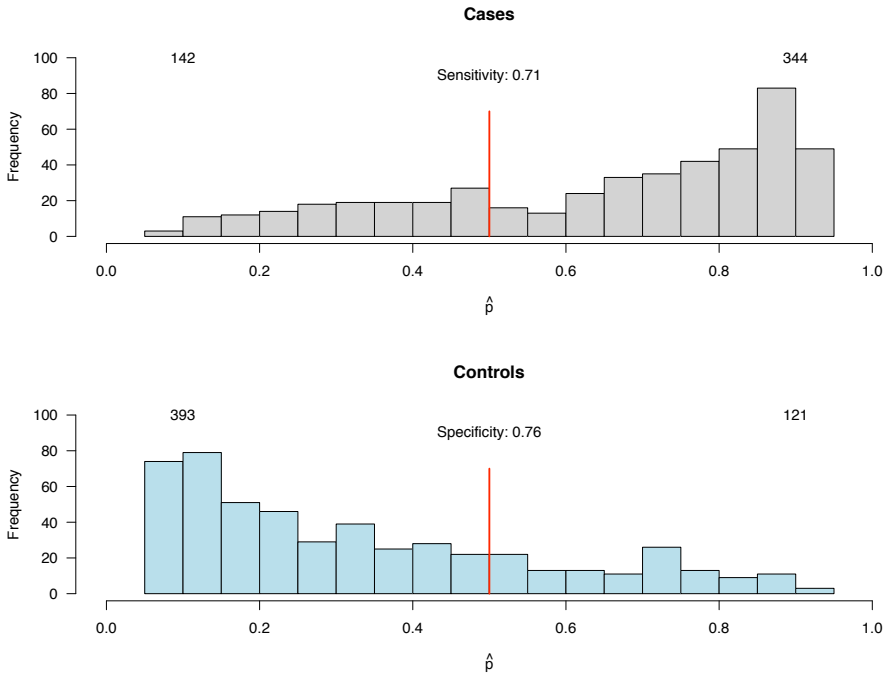
Fitted probabilities



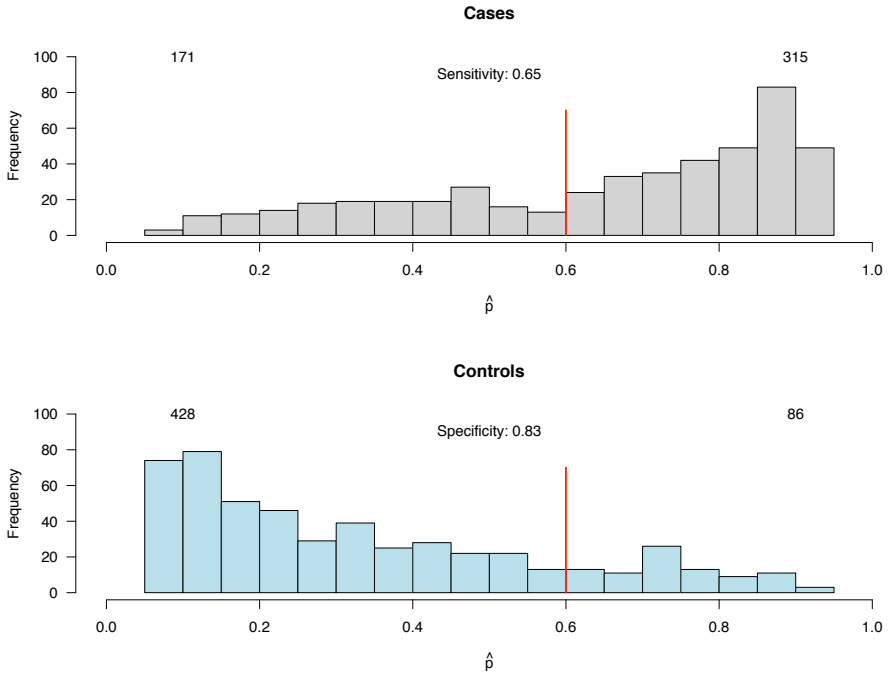
Fitted probabilities



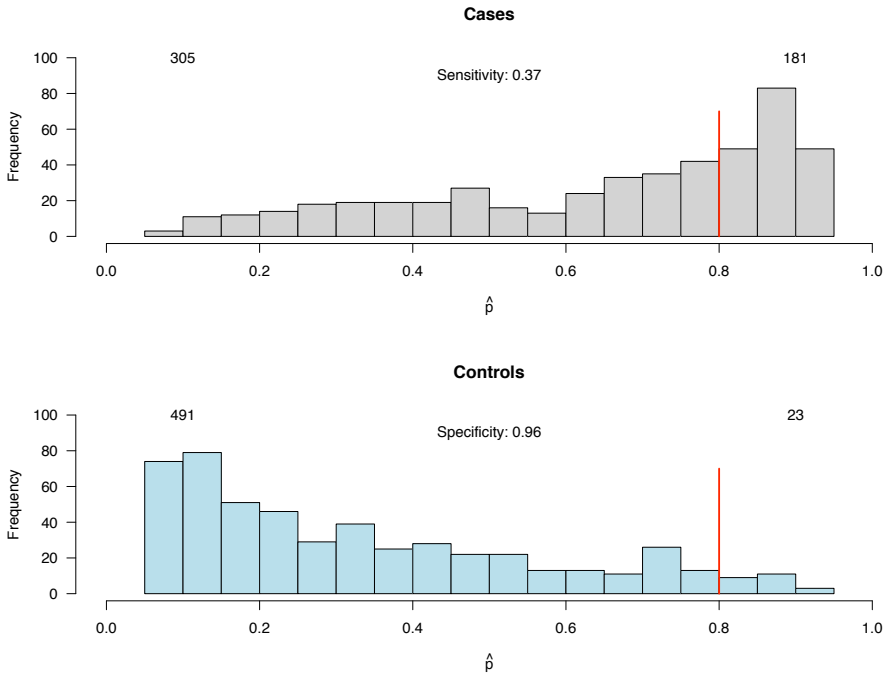
Fitted probabilities



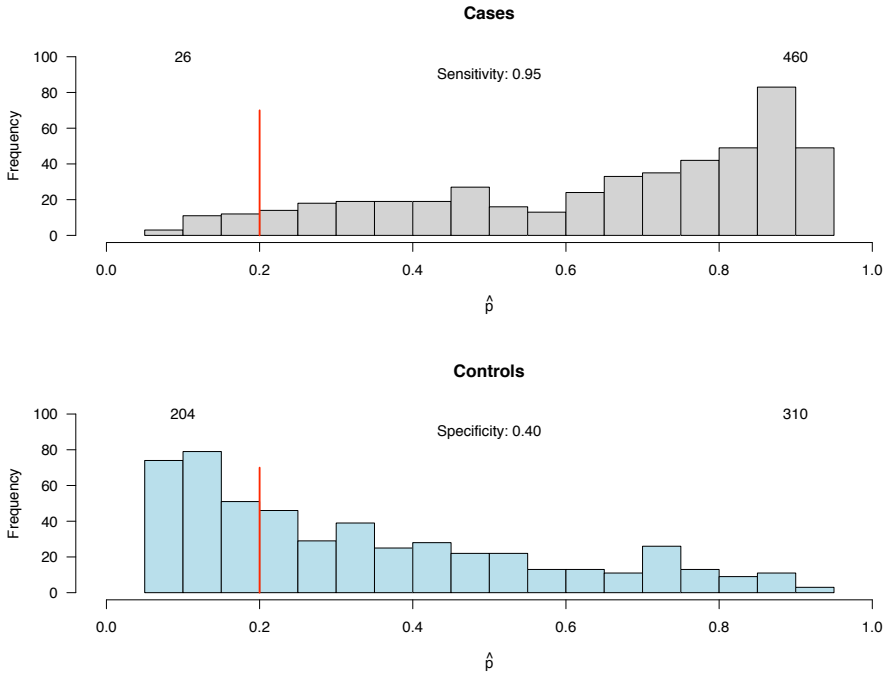
Fitted probabilities



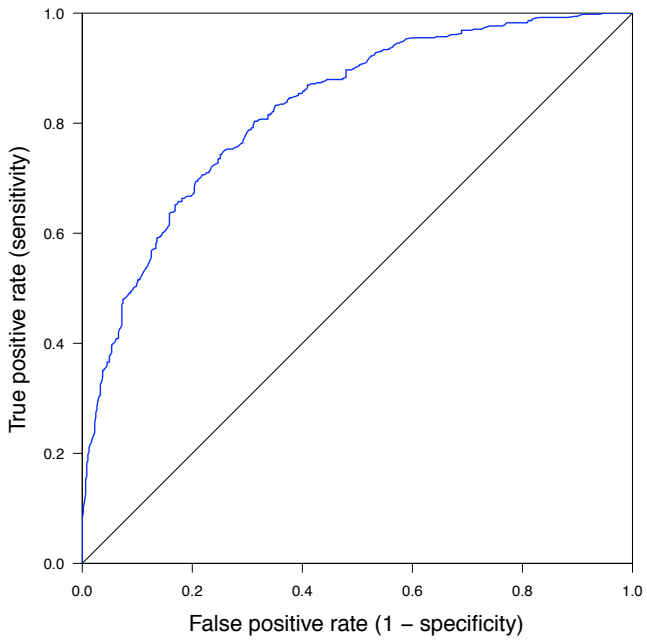
Fitted probabilities



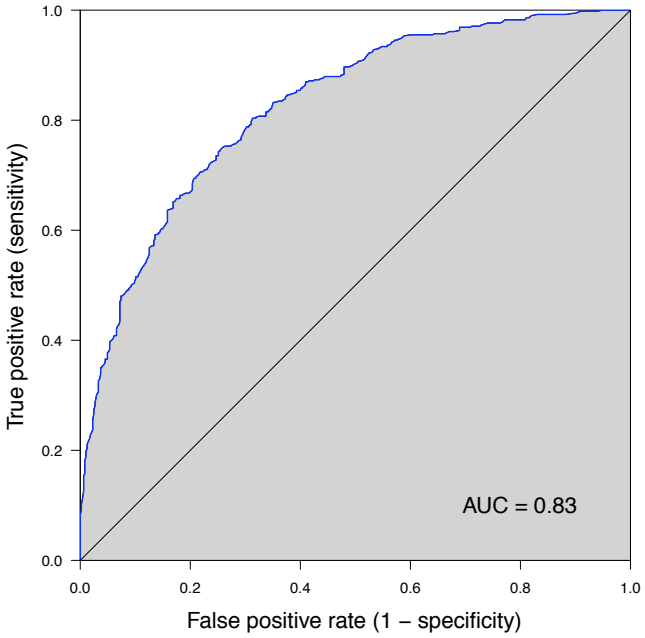
Fitted probabilities



ROC curve



ROC curve



Propensity scores

Suppose that a researcher wishes to compare the long-term survival of patients who received coronary artery bypass surgery with those who did not receive surgery. Patients selected for CABG can be expected to differ from those that did not receive surgery in terms of important prognostic characteristics including the severity of coronary artery disease or the presence of concurrent conditions, such as diabetes. A simple comparison of the survival of patients who either did or did not receive CABG will be biased by these confounding variables. This "confounding by indication" is almost invariably present in non-randomised studies of healthcare interventions and is difficult to overcome.

Nicholas J, Gulliford MC (2008)

Propensity scores

Rosenbaum and Rubin (1983) proposed the use of propensity scores as a method for allowing for confounding by indication. Propensity may be defined as an individual's probability of being treated with the intervention of interest given the complete set of all information about that individual. The propensity score provides a single metric that summarises all the information from explanatory variables such as disease severity and comorbity; it estimates the probability of a subject receiving the intervention of interest given his or her clinical status.

Nicholas J, Gulliford MC (2008)

Propensity scores

⟶ The propensity score is the conditional probability of receiving a given exposure (treatment), given a vector of measured covariates.

The propensity score is usually estimated via logistic regression.

Let T be the treatment and $X_1, \ldots, X_k$ be the covariates recorded.

$$\text{logit}(p(T)) = \beta_0 + \beta_1 \times X_1 + \cdots + \beta_k \times X_k.$$

The propensity score is the fitted probability of treatment, given the covariates.

⟶ The propensity score calculation does not use the outcome Y.

⟶ We have to assume that treatment assignment and the potential outcomes are conditionally independent.

Example

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	*P* Value
Demographics			
Age, mean (SD), y	62 (11)	56 (12)	<.001
Men, No. (%)	1779 (77)	2167 (56)	<.001
Clinical history			
Diabetes, No. (%)	388 (17)	432 (11)	<.001
Hypertension, No. (%)	1224 (53)	1569 (41)	<.001
Tobacco use, No. (%)	234 (10)	500 (13)	.001
Prior coronary artery disease, No. (%)	1609 (70)	778 (20)	<.001
Prior coronary artery bypass graft, No. (%)	689 (30)	240 (6)	<.001
Prior percutaneous coronary intervention, No. (%)	667 (29)	148 (4)	<.001
Prior Q-wave MI, No. (%)	369 (16)	285 (7)	<.001
Atrial fibrillation, No. (%)	27 (1)	55 (1)	.04
Congestive heart failure, No. (%)	127 (6)	178 (5)	.12
Medication use			
Digoxin use, No. (%)	171 (7)	216 (6)	.004
β-Blocker use, No. (%)	811 (35)	550 (14)	<.001
Diltiazem/verapamil use, No. (%)	452 (20)	405 (10)	<.001
Nifedipine use, No. (%)	261 (11)	283 (7)	<.001
Lipid-lowering therapy, No. (%)	775 (34)	380 (10)	<.001
ACE inhibitor use, No. (%)	349 (15)	441 (11)	<.001

Gum et al (2001)

Example

Table 3. Selected Baseline and Exercise Characteristics According to Aspirin Use in Propensity-Matched Patients*

Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	*P* Value
Demographics			
Age, mean (SD), y	60 (11)	61 (11)	.16
Men, No. (%)	951 (70)	974 (72)	.33
Clinical history			
Diabetes, No. (%)	203 (15)	207 (15)	.83
Hypertension, No. (%)	679 (50)	698 (52)	.46
Tobacco use, No. (%)	161 (12)	162 (12)	.95
Cardiac variables			
Prior coronary artery disease, No. (%)	652 (48)	659 (49)	.79
Prior coronary artery bypass graft, No. (%)	251 (19)	235 (17)	.42
Prior percutaneous coronary intervention, No. (%)	166 (12)	147 (11)	.25
Prior Q-wave MI, No. (%)	194 (14)	206 (15)	.52
Atrial fibrillation, No. (%)	21 (2)	24 (2)	.65
Congestive heart failure, No. (%)	79 (6)	89 (7)	.43
Medication use			
Digoxin use, No. (%)	115 (9)	114 (9)	.94
β-Blocker use, No (%)	352 (26)	358 (26)	.79
Diltiazem/verapamil use, No. (%)	223 (17)	223 (17)	>.99
Nifedipine use, No. (%)	127 (9)	144 (11)	.28
Lipid-lowering therapy, No. (%)	281 (21)	271 (20)	.63
ACE inhibitor use, No. (%)	209 (15)	214 (16)	.79

Gum et al (2001)

Log-linear models

Higher order contingency tables are frequently analysed using log-linear models. The below is a tabulation of breast cancer data from Morrison et al. Recorded were diagnostic center, nuclear grade, and survival.

	malignant died	malignant survived	benign died	benign survived
Boston	35	59	47	112
Glamorgan	42	77	26	76

$$\log(\hat{f}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

$\longrightarrow$ We are mostly interested in the interactions!

Log-linear models

The saturated model:

variable	p-value
(Intercept)	0.00
center	0.42
grade	0.18
survival	0.01
center $\times$ grade	0.02
center $\times$ survival	0.76
grade $\times$ survival	0.20
grade $\times$ center $\times$ survival	0.76

Log-linear models

A sub-model:

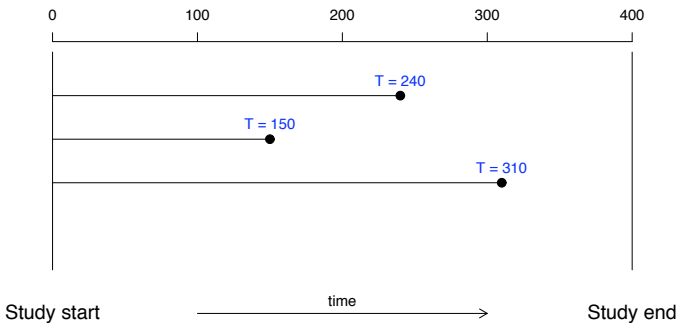
variable	p-value
(Intercept)	0.00
center	0.08
grade	0.15
survival	0.00
center $\times$ grade	0.00
grade $\times$ survival	0.05

Survival Analysis

Survival analysis

Survival analysis: Study of durations between events

⟶ Outcome:
Time until an event occurs, i.e. *survival time* or *failure time*.

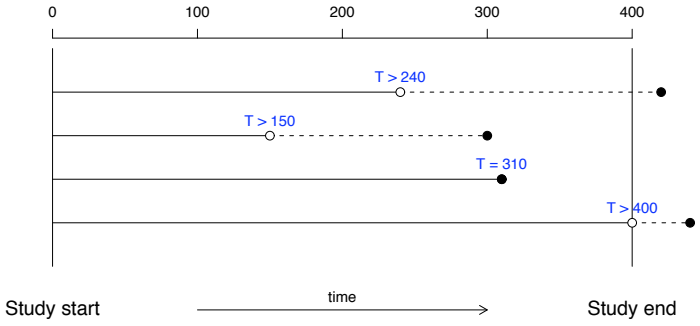


Examples: Age at death, age at first disease diagnosis, waiting time to pregnancy, duration between treatment and death, . . .

The censoring problem in survival analysis

⟶ Censoring:
Incomplete observations of the survival time.

⟶ Right censoring:
Some individuals may not be observed for the full time to failure, because of loss to follow-up, drop out, termination of the study, . . .



Basic goals of survival analysis

1. To estimate and interpret survival characteristics
   $\longrightarrow$ Kaplan-Meier plots

2. To compare survival in different groups
   $\longrightarrow$ Log-rank test

3. To assess the relationship of explanatory variables to survival
   $\longrightarrow$ Cox regression model

Survival function

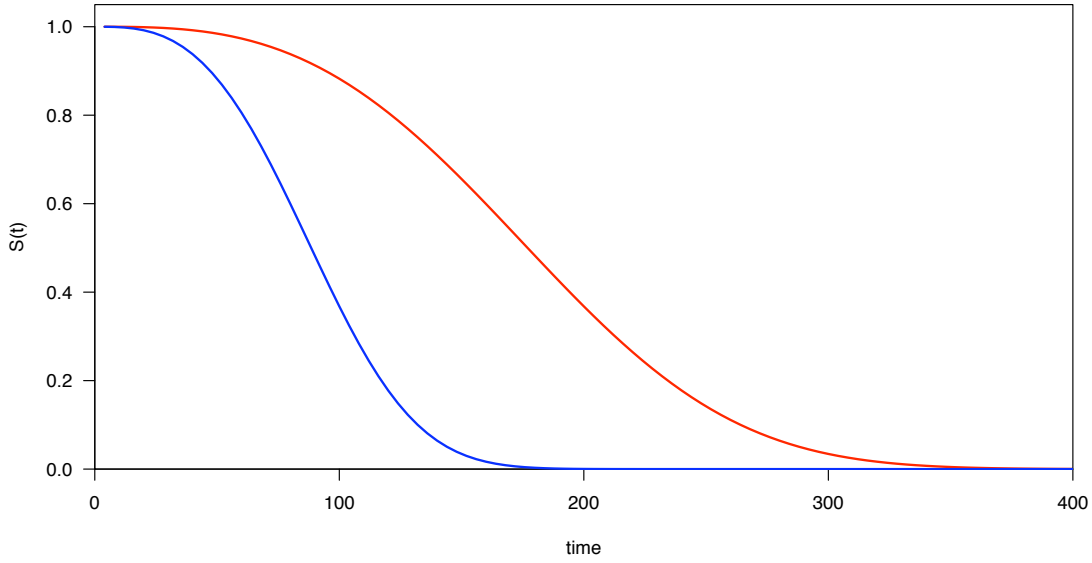
Survival function: $S(t) = P(T > t)$

$\longrightarrow$ S(t) describes the probability of surviving to time t, or what fraction of subjects survive (on average) to time t.

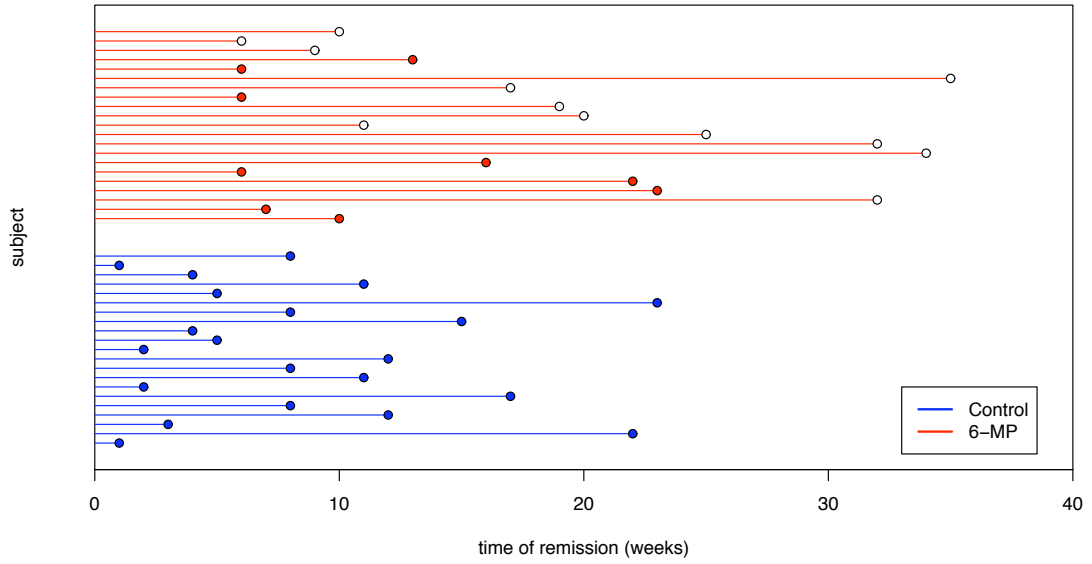
Properties:

- S(t) is a smooth function in t.
- $S(0) = 1$ and $S(\infty) = 0$.
- S(t) is a decreasing function in t.
- Describes *cumulative* survival characteristics.

Survival functions



Example



Kaplan-Meier estimate

The Kaplan-Meier or product-limit estimate $\hat{S}(t)$ is an estimate of $S(t)$ from a finite sample.

Suppose that there are observations on n individuals and assume that there are k ($k \leq n$) distinct times $t_1, \ldots, t_k$ at which deaths occur. Let d_j be the number of deaths at time t_j. Define

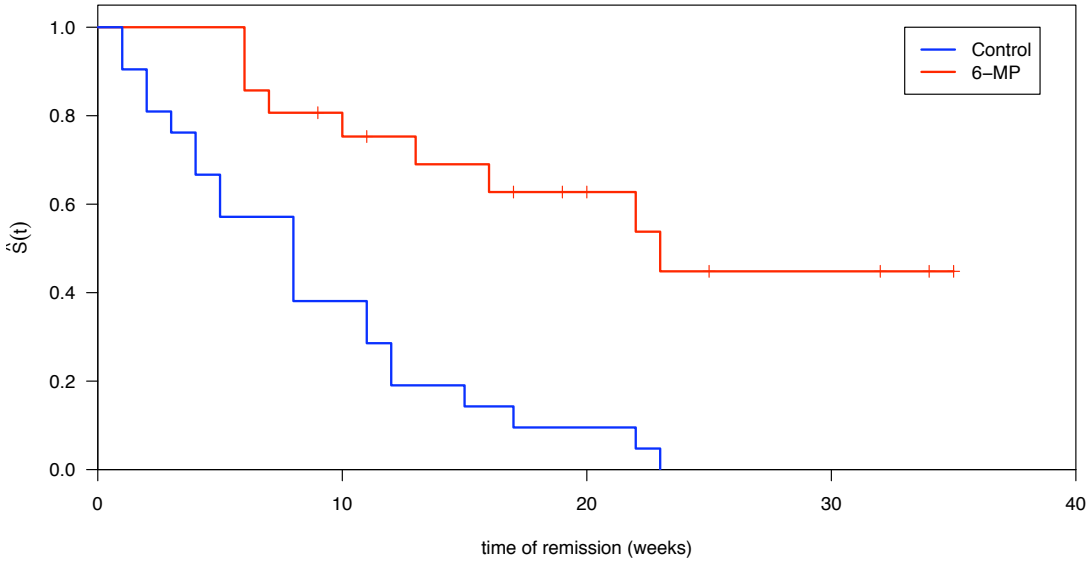
$$\hat{S}(t) = \prod_{j:\ t_j<t} \frac{n_j - d_j}{n_j},$$

where n_j is the number of individuals at risk (e.g., the individuals alive and uncensored) at time t_j.

$\longrightarrow$ If there are no censored observations, this reduces to

$$\hat{S}(t) = \text{(number of observations} \geq t) \text{ / n.}$$

Example



Some facts about the Kaplan-Meier estimate

⟶ The Kaplan-Meier method is *non-parametric*. The survival curve is step-wise, not smooth. Any jumping point is a failure time point. The jump size is proportional to the number of deaths at a failure time point. Note that having a small sample means having big steps!

⟶ If the largest observed study time t_k corresponds to a death time, then the estimated Kaplan-Meier survival curve is 0 beyond t_k. If the largest observed study time is censored, then the survival curve is not 0 beyond t_k.

⟶ $\hat{S}(t)$ is a decreasing function in t with $\hat{S}(0) = 1$. Further $\hat{S}(t)$ converges to S(t) as $n \to \infty$.

Comparison of two survival distributions

We test H_0: $S_1(t) = S_2(t)$ versus H_a: $S_1(t) \neq S_2(t)$

⟶ The main idea behind the two-sample log-rank test: if survival is unrelated to group effect, then at each time point, roughly the same proportion in each group will fail.

The test is based on χ^2-types of statistics:

$$Q = \sum_{i=1}^{D} (O_{1i} - E_{1i})$$

where the summation is over the pooled failure time points among the 2 groups. O_{1i} and E_{1i} are the observed number of death for group 1 at the i^{th} pooled failure time. The log-rank test statistic under H_0 is

$$\text{logRT} = \frac{Q^2}{\text{Var}(Q)} \sim \chi^2_1$$

Example

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP     21        9     19.3      5.46      16.8
treat=control  21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Comparison of survival distributions

The log-rank test can be extended to $k > 2$ groups. Under H_0 the null distribution of the test statistic is

$$\text{logRT} \sim \chi^2_{k-1}$$

However, these test also have some shortcomings:

- The tests have a bad performance when the two survival functions are overcrossing.
- The test can only be used for comparing groups defined by single categorical covariates.
- They are not very useful to quantify the differences.

Hazard function

The hazard function is defined as

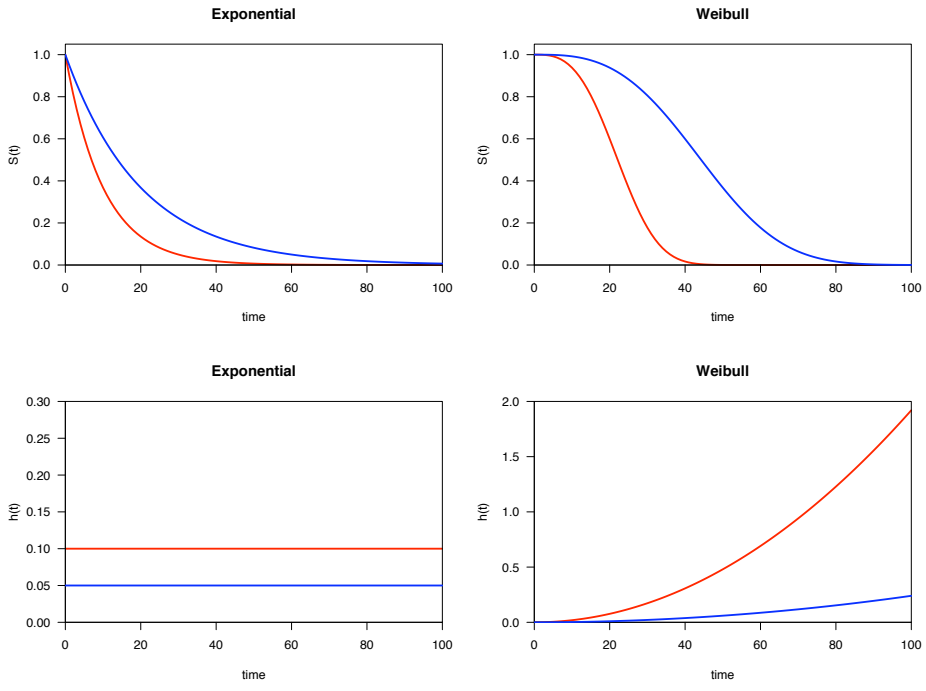
$$h(t) = -\frac{d}{dt}\log(S(t))$$

In other words, it is the slope of – log(S(t)). You can think of it as the propensity for failure for an individual at each time point, e.g. the instantaneous risk of failure.

Properties:

- Closely related to the incidence rate.
- Not a probability!
- May increase or decrease or both.
- Describes *instantaneous* survival characteristics.

Hazard functions



Cox regression model

⟶ Goal:

To assess the relationship of explanatory variables (e.g. sex, age, treatment type, etc) to survival time.

⟶ One idea (Sir David Cox):

Use a proportional hazards regression model, defined as

$$h(t|x) = h_0(t)e^{\beta x}$$

Here, $h_0(t)$ is a baseline hazard function, and β is a regression coefficient.

Cox regression model

What does $h(t|x) = h_0(t)e^{\beta x}$ mean?

For example, assume we a treatment group ($x = 1$) and a control group ($x = 0$).

⟶ In the control group, the hazard function is

$$h(t|x = 0) \;=\; h_0(t)e^{\beta \times 0} \;=\; h_0(t)$$

⟶ In the treatment group, the hazard function is

$$h(t|x = 1) \;=\; h_0(t)e^{\beta \times 1} \;=\; h_0(t)e^{\beta}$$

⟶ The relative risk for treatment versus control group is

$$\text{RR} \;=\; \frac{h(t|x = 1)}{h(t|x = 0)} \;=\; e^{\beta}$$

Cox regression model

⟶ Interpretation of the parameters:

$\beta > 0$	RR > 1	and	$h(t\|x=1) > h(t\|x=0)$
$\beta = 0$	RR $= 1$	and	$h(t\|x=1) = h(t\|x=0)$
$\beta < 0$	RR < 1	and	$h(t\|x=1) < h(t\|x=0)$

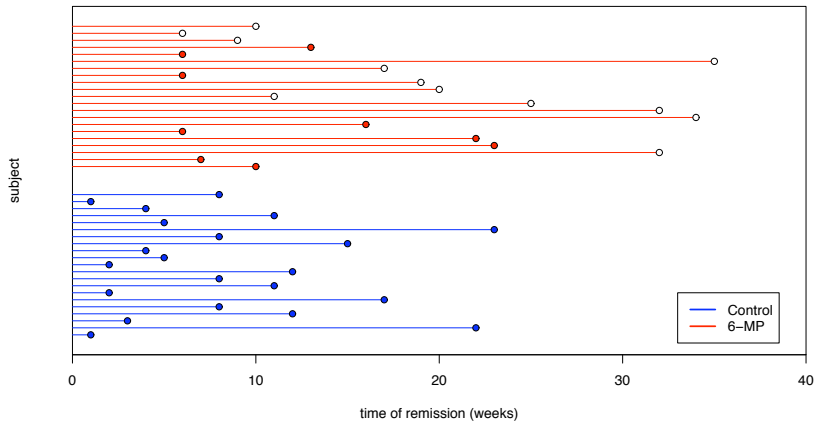
⟶ Hypothesis of interest:

$H_0 : \beta = 0$ (no treatment effect)

$H_a : \beta \neq 0$ (treatment influences survival)

Example

```
             coef exp(coef) se(coef)    z       p
treatcontrol 1.57      4.82    0.412 3.81 0.00014

             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol      4.82       0.208      2.15      10.8
```



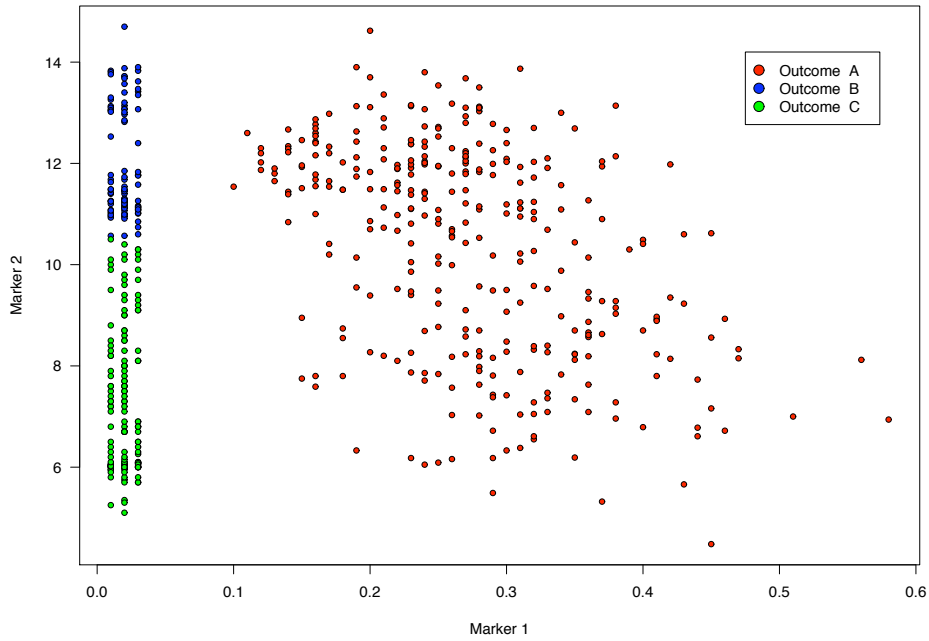
Another example

- Survival times for 33 patients who died from acute myelogenous leukaemia.
- Also measured was the patient's white blood cell count at the time of diagnosis.
- The patients were also factored into 2 groups according to the presence or absence of a morphologic characteristic of white blood cells (identified by the presence of Auer rods and/or significant granulation of the leukaemic cells in the bone marrow at the time of diagnosis).

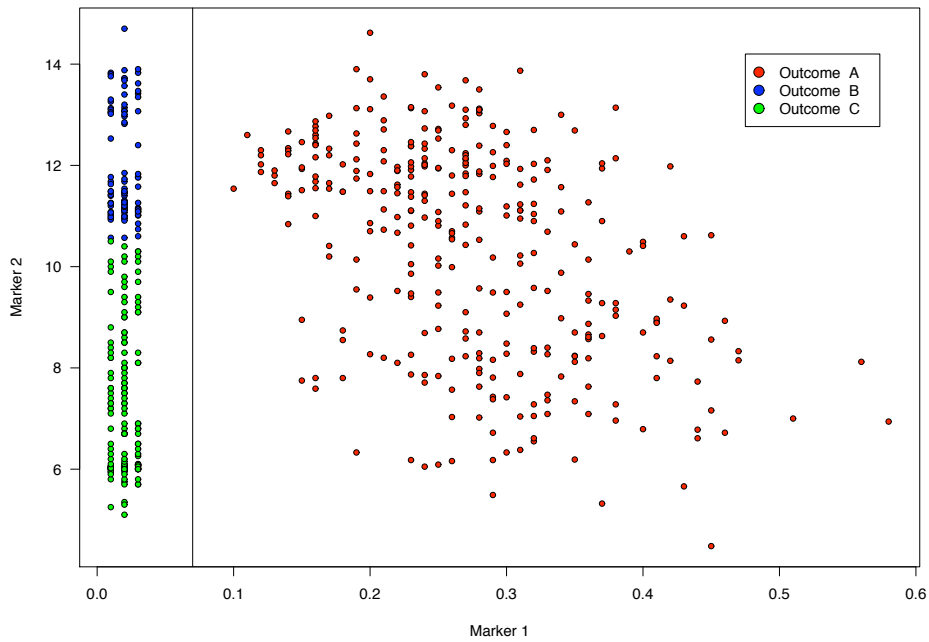
```
             coef exp(coef) se(coef)     z      p
agpresent -1.069     0.343    0.429 -2.49 0.0130
log(wbc)   0.368     1.444    0.136  2.70 0.0069

          exp(coef) exp(-coef) lower .95 upper .95
agpresent     0.343      2.913     0.148     0.796
log(wbc)      1.444      0.692     1.106     1.886
```

Classification and Regression Trees

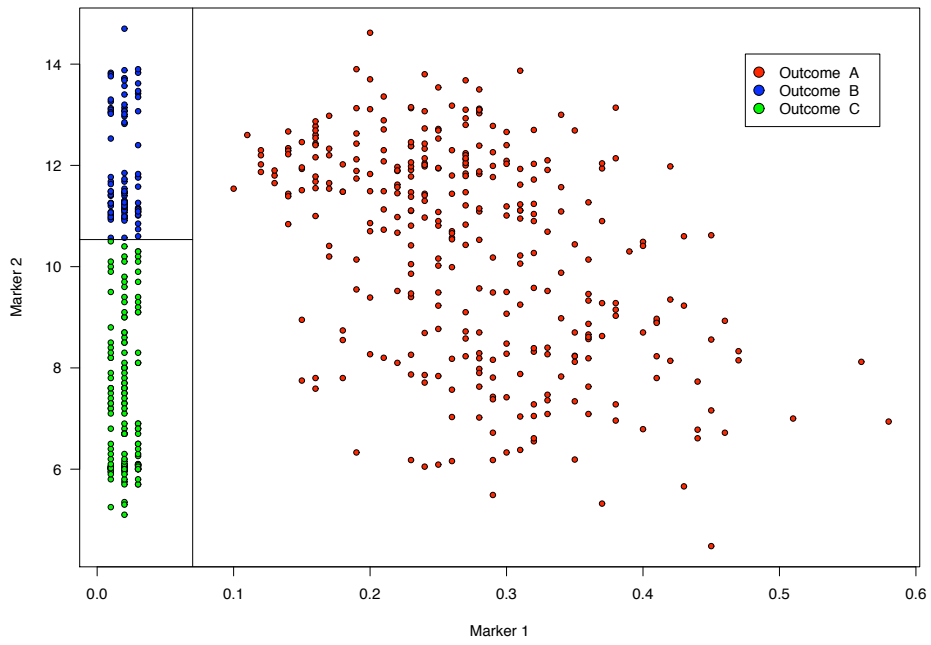
Example 1



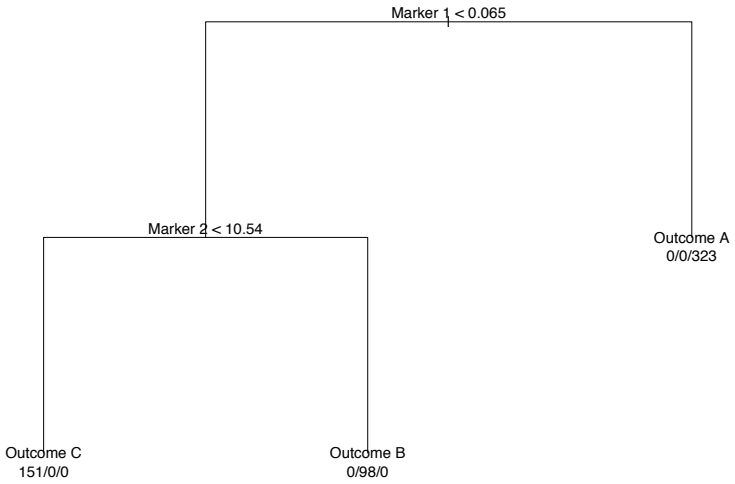
Example 1



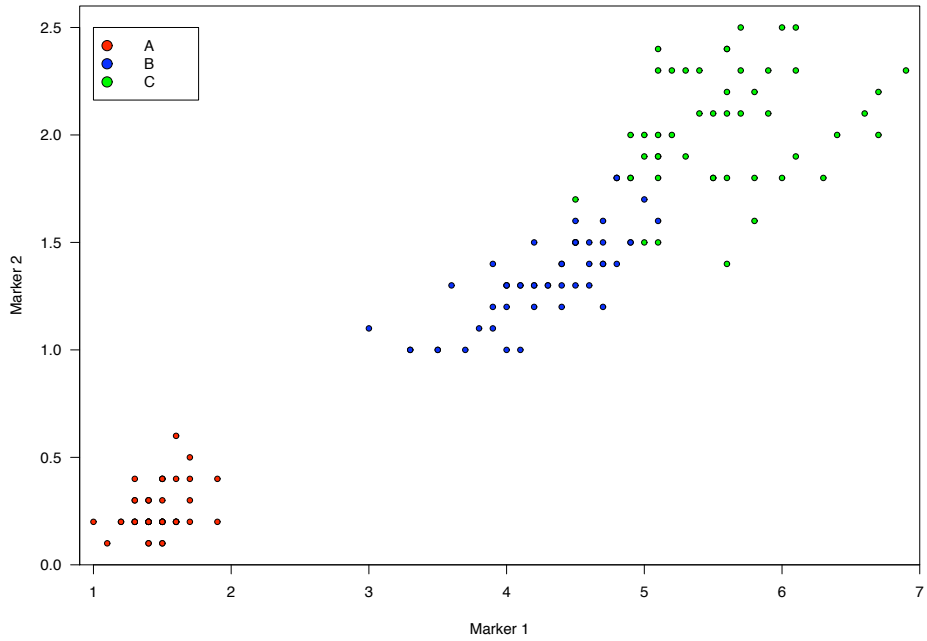
Example 1



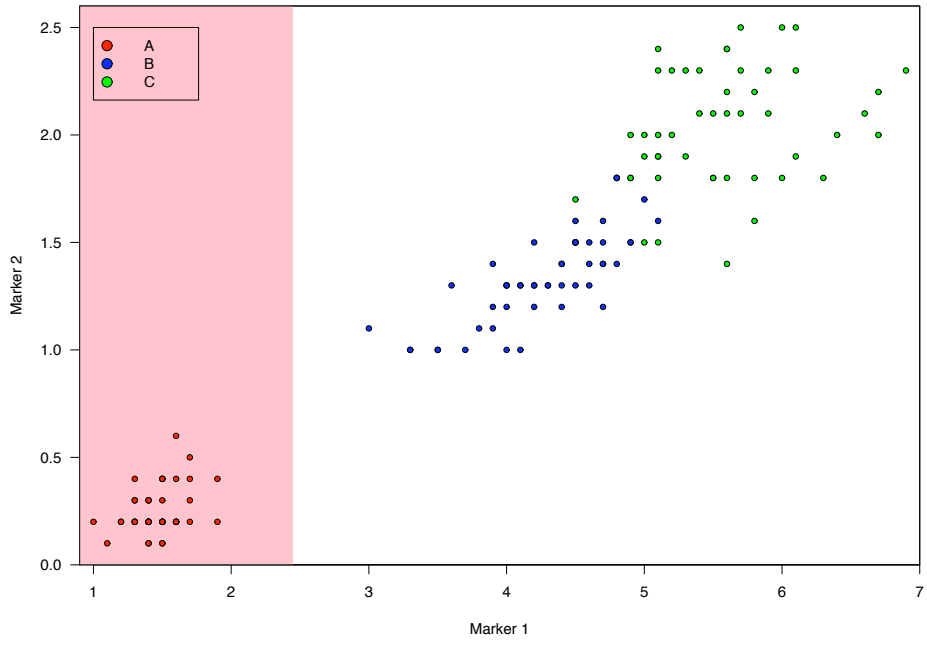
Example 1



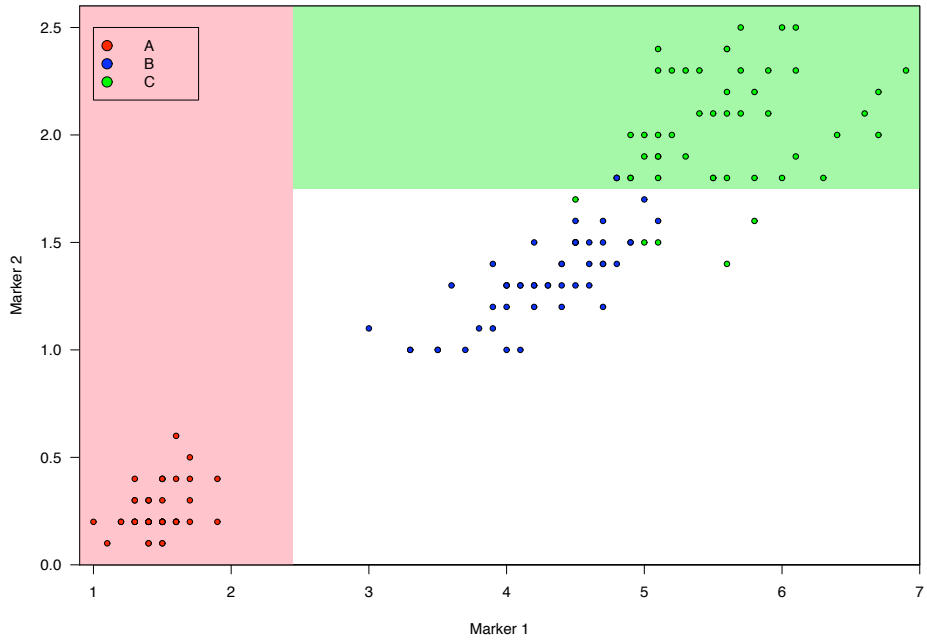
Example 2



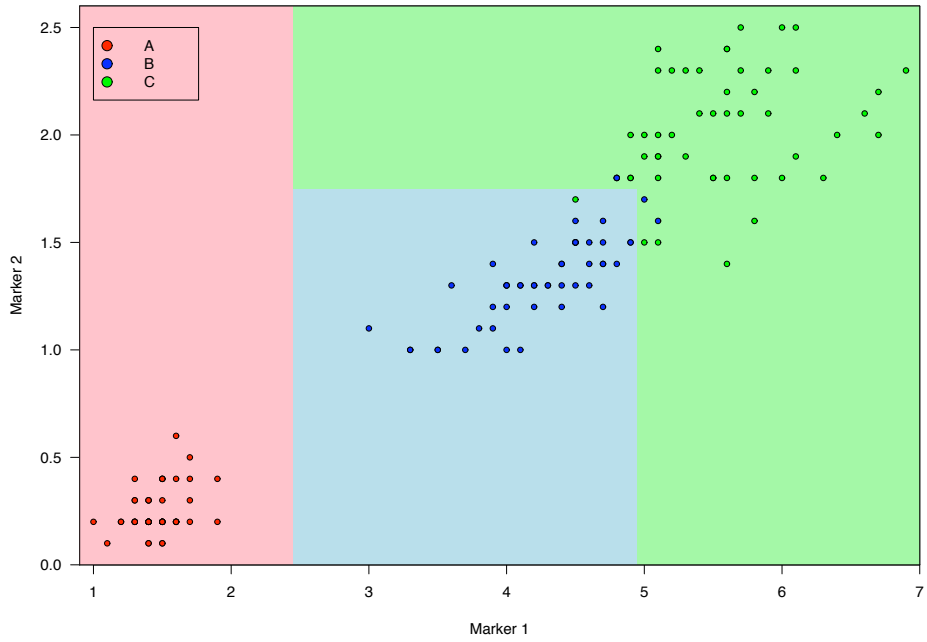
Example 2



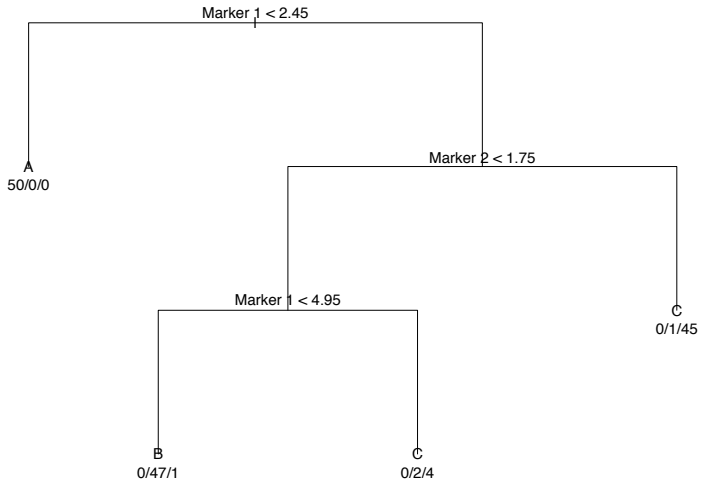
Example 2



Example 2

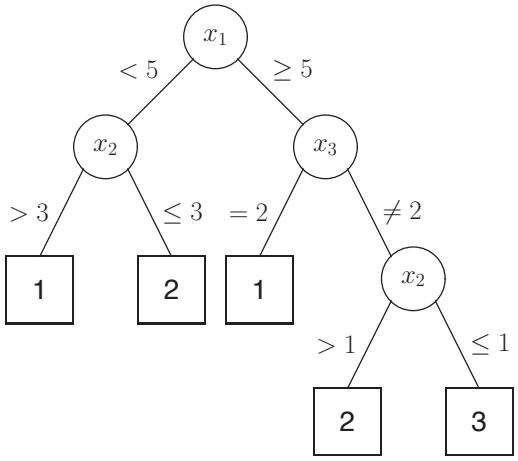


Example 2



Classification Tree

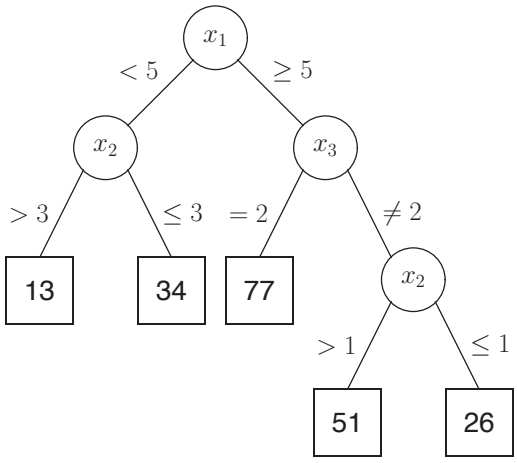
Suppose that we have a scalar outcome, Y, and a p-vector of explanatory variables, X. Assume $Y \in \mathcal{K} = \{1, 2, \ldots, k\}$



A classification tree partitions the X-space and provides a predicted value, perhaps $\arg\max_s \Pr(Y = s | X \in A_k)$ in each region.

Regression Tree

Again, suppose that we have a scalar outcome, Y, and a p-vector of explanatory variables, X. Now assume $Y \in \mathcal{R}$.



A regression tree partitions the X-space into disjoint regions A_k and provides a fitted value $\mathrm{E}(Y|X \in A_k)$ within each region.

Recursive Partitioning

INITIALIZE	All cases in the root node.
REPEAT	Find optimal allowed split.
	Partition leaf according to split.
STOP	Stop when pre-defined criterion is met.

The Predictor Space

Suppose that we have p explanatory variables $X_1, \ldots, X_p$ and n observations.

Each of the X_i can be

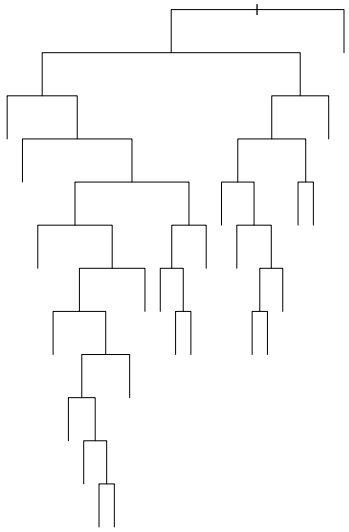
a) a numeric variable:
$\longrightarrow$ $n-1$ possible splits.

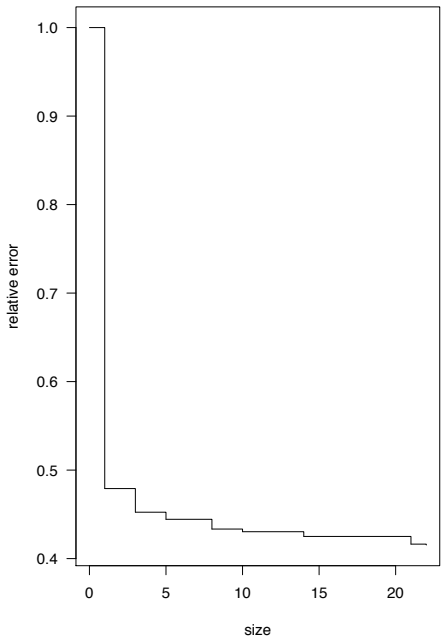
b) an ordered factor:
$\longrightarrow$ $k-1$ possible splits.

b) an unordered factor:
$\longrightarrow$ $2^{k-1}-1$ possible splits.

We pick the split that results in the greatest decrease in impurity (according to some impurity measure).

Trees





Example: Low Birth Weight Data

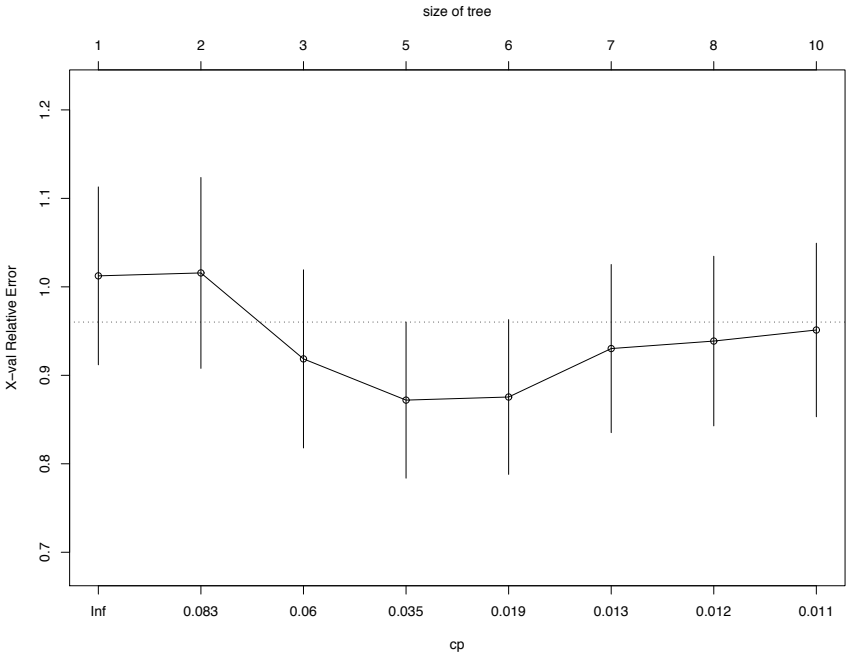
Problem: Predict a child's birthweight from a list of variables.

The birth weight data were collected in 1986 at the Baystate Medical Center, Springfield, MA. For 189 infants, the following variables are available:

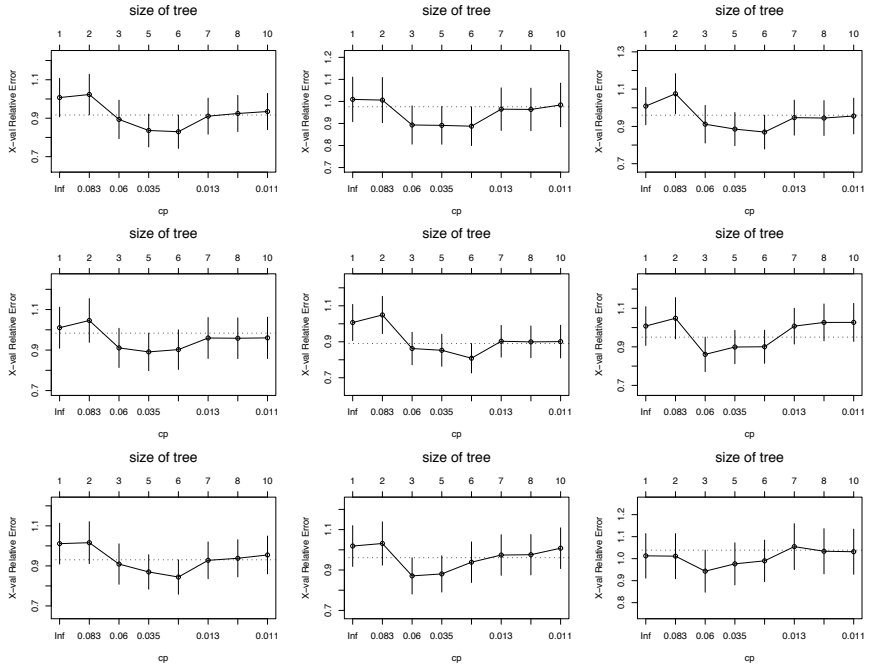
- an indicator of birth weight less than 2500g (yes/no),
- the mother's age in years,
- the mother's weight in pounds at last menstrual period,
- the mother's race (white/black/other),
- the smoking status during pregnancy (yes/no),
- the number of previous premature labours,
- the history of hypertension (yes,no),
- the presence of uterine irritability (yes/no),
- the number of physician visits during the first trimester,
- the birth weight (grams).

Reference: Hosmer, DW and Lemeshow, S (1989). *Applied Logistic Regression*, New York: Wiley.

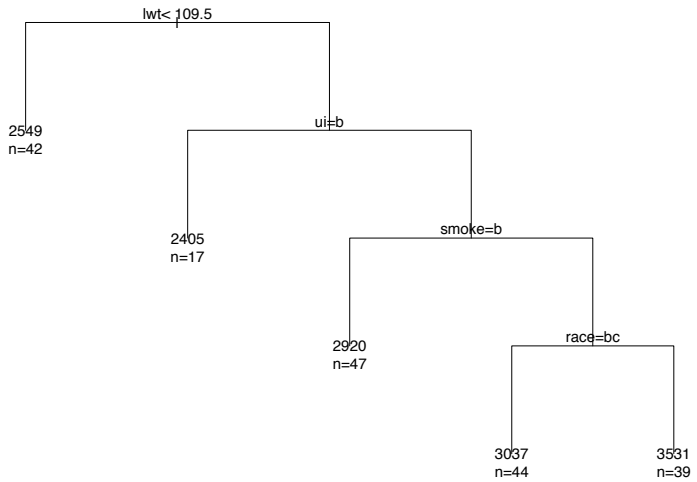
Example: Low Birth Weight Data



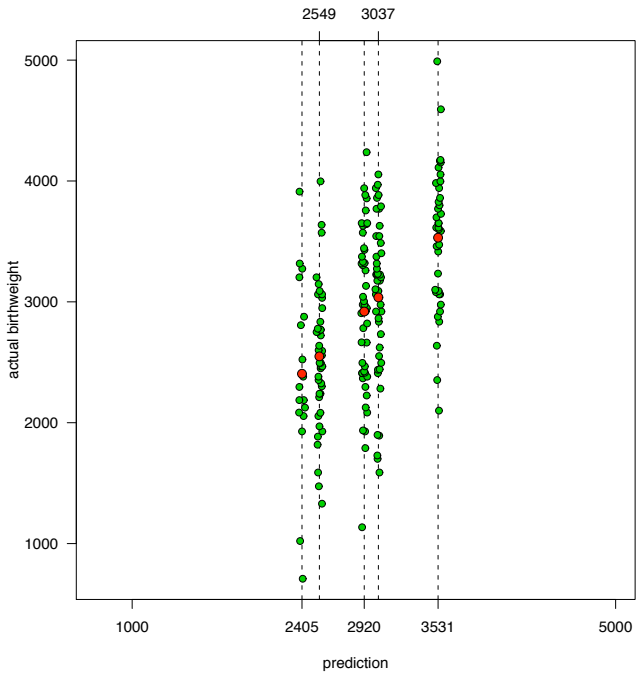
Example: Low Birth Weight Data



Example: Low Birth Weight Data



Example: Low Birth Weight Data



General Points

What's nice:

- Decision trees are very "natural" constructs, in particular when the explanatory variables are categorical (and even better, when they are binary).
- Trees are very easy to explain and interpret.
- The models are invariant under transformations in the predictor space.
- Multi-factor response is easily dealt with.
- The treatment of missing values is more satisfactory than for most other model classes.
- The models go after interactions immediately, rather than as an afterthought.
- The tree growth is actually more efficient than I have described it.
- There are extensions for survival and longitudinal data, and there is an extension called treed models. There is even a Bayesian version of CART.

General Points

What's not so nice:

- The tree-space is huge, so we may need a lot of data.
- We might not be able to find the "best" model at all.
- It can be hard to assess uncertainty in inference about trees.
- The results can be quite variable (the tree selection is not very stable).
- Actual additivity becomes a mess in a binary tree.
- Simple trees usually do not have a lot of predictive power.
- There is a selection bias for the splits.

Other supervised approaches

- Bagging
- Random forests
- Support vector machines
- Linear discriminant analysis
- . . .