Logistic Regression

Example

<table>
<thead>
<tr>
<th>Dose</th>
<th>No. survived</th>
<th>No. dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>0.5</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>1.0</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>1.5</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>2.0</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>3.0</td>
<td>1</td>
<td>24</td>
</tr>
</tbody>
</table>
Binary vs. continuous outcomes

Continuous: ANOVA ←→ Regression

Binary: $k \times 2$ table ←→ ?

Goals:

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

$\rightarrow$ 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:

\[ \mathbb{E}(y \mid x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \]

\[ \text{What is the interpretation of } \beta_i ? \]

Binary outcomes

Let \( p_d = \Pr(\text{dead} \mid \text{dose } d) \)

\[ p_d = \beta_0 + \beta_1 d \]

\[ 0 \leq p_d \leq 1 \quad \text{but} \quad -\infty \leq \beta_0 + \beta_1 d \leq \infty \]

Odds of death:

\[ 0 \leq \frac{p_d}{1 - p_d} \leq \infty \]

Log odds of death:

\[ -\infty \leq \ln \left( \frac{p_d}{1 - p_d} \right) \leq \infty \]

\[ \ln \left( \frac{p}{1 - p} \right) \text{ is also called logit}(p) \text{ or the logistic function.} \]
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:

\[\rightarrow \] What if \( \hat{p}_d = 0 \) or 1?

\[\rightarrow \] 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,
  - \( \logit(p_d) = \ln\left(\frac{p_d}{1-p_d}\right) = \beta_0 + \beta_1 d \)

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

Likelihood:
\[
L(\beta_0, \beta_1 | y) = \prod_d p_d^{y_d} (1 - p_d)^{(n_d - y_d)}
\]

Logistic regression

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

Software output:

> summary(glm.out)$coef

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
<th>t-val</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.33</td>
<td>0.33</td>
<td>-4.06</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>dose</td>
<td>1.44</td>
<td>0.23</td>
<td>6.29</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Interpretation of $\beta$’s

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

$\beta_0 = \log$ odds when dose = 0

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

$\beta_1 = \text{change in log odds with unit increase in dose}$

Note: $\beta_1 = 0 \quad \rightarrow \quad \text{survival unrelated to dose.}$
**LD50**

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

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

\[0 = \beta_0 + \beta_1 (LD50)\]

\[LD50 = -\frac{\hat{\beta}_0}{\hat{\beta}_1}\]

\[\hat{LD50} = -\frac{\hat{\beta}_0}{\hat{\beta}_1}\]

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

<table>
<thead>
<tr>
<th>Sex</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Female</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

A plot of the data
Analysis

Assume no sex difference

\[ \logit(p) = \beta_0 + \beta_1 \times \text{dose} \]

\[
> \text{summary(glm.out)$coef}
\]

<table>
<thead>
<tr>
<th>Est</th>
<th>SE</th>
<th>t-val</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.57</td>
<td>0.23</td>
<td>-6.8</td>
</tr>
<tr>
<td>dose</td>
<td>0.153</td>
<td>0.022</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Assume sexes completely different

\[ \logit(p) = \beta_0 + \beta_1 \times \text{sex} + \beta_2 \times \text{dose} + \beta_3 \times \text{sex:dose} \]

\[
> \text{summary(glm.out)$coef}
\]

<table>
<thead>
<tr>
<th>Est</th>
<th>SE</th>
<th>t-val</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.72</td>
<td>0.32</td>
<td>-5.3</td>
</tr>
<tr>
<td>sexmale</td>
<td>-0.21</td>
<td>0.51</td>
<td>-0.4</td>
</tr>
<tr>
<td>dose</td>
<td>0.116</td>
<td>0.024</td>
<td>4.9</td>
</tr>
<tr>
<td>sexmale:dose</td>
<td>0.182</td>
<td>0.067</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Analysis (continued)

Different slopes but common “intercept”

\[ \logit(p) = \beta_0 + \beta_1 \times \text{dose} + \beta_2 \times \text{sex:dose} \]

\[
> \text{summary(glm.out)$coef}
\]

<table>
<thead>
<tr>
<th>Est</th>
<th>SE</th>
<th>t-val</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.80</td>
<td>0.25</td>
<td>-7.2</td>
</tr>
<tr>
<td>dose</td>
<td>0.120</td>
<td>0.021</td>
<td>5.6</td>
</tr>
<tr>
<td>dose:sexmale</td>
<td>0.161</td>
<td>0.044</td>
<td>3.7</td>
</tr>
</tbody>
</table>
Fitted curves

Plot using \( \log_2 \) dose
Use log\(_2\) of the dose

Assume no sex difference

\[ \text{logit}(p) = \beta_0 + \beta_1 \times \log_2(\text{dose}) \]

```r
> 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

\[ \text{logit}(p) = \beta_0 + \beta_1 \times \text{sex} + \beta_2 \times \log_2(\text{dose}) + \beta_3 \times \text{sex:log}_2(\text{dose}) \]

```r
> 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”

\[ \text{logit}(p) = \beta_0 + \beta_1 \times \log_2(\text{dose}) + \beta_2 \times \text{sex:log}_2(\text{dose}) \]

```r
> 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

Cases

Controls

Fitted probabilities

Cases

Controls

Sensitivity: 0.71

Specificity: 0.76
Fitted probabilities

Cases

Frequency

Sensitivity: 0.65

Controls

Specificity: 0.83

Fitted probabilities

Cases

Frequency

Sensitivity: 0.37

Controls

Specificity: 0.96
Fitted probabilities

Cases

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Sensitivity: 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Controls

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Specificity: 0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ROC curve

0.0 0.2 0.4 0.6 0.8 1.0

False positive rate (1 – specificity)

True positive rate (sensitivity)
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.
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 comorbidity; it estimates the probability of a subject receiving the intervention of interest given his or her clinical status.

Nicholas J, Gulliford MC (2008)

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 calculation does not use the outcome $Y$.

We have to assume that treatment assignment and the potential outcomes are conditionally independent.
### Table 1. Baseline and Exercise Characteristics According to Aspirin Use

<table>
<thead>
<tr>
<th>Variable</th>
<th>Aspirin (n = 2310)</th>
<th>No Aspirin (n = 3604)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, mean (SD), y</td>
<td>62 (11)</td>
<td>55 (12)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Men, No. (%)</td>
<td>1779 (77)</td>
<td>2167 (86)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Clinical History</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes, No. (%)</td>
<td>361 (17)</td>
<td>432 (11)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hypertension, No. (%)</td>
<td>1224 (53)</td>
<td>1569 (41)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Tobacco use, No. (%)</td>
<td>234 (10)</td>
<td>500 (12)</td>
<td>.001</td>
</tr>
<tr>
<td>Prior coronary artery disease, No. (%)</td>
<td>1009 (70)</td>
<td>778 (62)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Prior coronary artery bypass graft, No. (%)</td>
<td>689 (60)</td>
<td>240 (9)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Prior percutaneous coronary intervention, No. (%)</td>
<td>867 (35)</td>
<td>148 (4)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Prior O-wave M, No. (%)</td>
<td>319 (15)</td>
<td>286 (7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Atrial fibrillation, No. (%)</td>
<td>27 (1)</td>
<td>55 (1)</td>
<td>.04</td>
</tr>
<tr>
<td>Congestive heart failure, No. (%)</td>
<td>127 (6)</td>
<td>176 (5)</td>
<td>.12</td>
</tr>
<tr>
<td><strong>Medication use</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digoxin use, No. (%)</td>
<td>171 (7)</td>
<td>216 (9)</td>
<td>.004</td>
</tr>
<tr>
<td>β-Blocker use, No. (%)</td>
<td>811 (36)</td>
<td>550 (14)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diltiazem/verapamil use, No. (%)</td>
<td>452 (20)</td>
<td>405 (12)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nifedipine use, No. (%)</td>
<td>251 (11)</td>
<td>283 (7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Lipid-lowering therapy, No. (%)</td>
<td>775 (34)</td>
<td>350 (10)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ACE inhibitor use, No. (%)</td>
<td>349 (15)</td>
<td>441 (11)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Gum et al (2001)

### Table 3. Selected Baseline and Exercise Characteristics According to Aspirin Use in Postmenopausal Patients

<table>
<thead>
<tr>
<th>Variable</th>
<th>Aspirin (n = 1351)</th>
<th>No Aspirin (n = 1351)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, mean (SD), y</td>
<td>60 (11)</td>
<td>61 (11)</td>
<td>.16</td>
</tr>
<tr>
<td>Men, No. (%)</td>
<td>951 (70)</td>
<td>914 (73)</td>
<td>.33</td>
</tr>
<tr>
<td><strong>Clinical History</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes, No. (%)</td>
<td>202 (15)</td>
<td>207 (15)</td>
<td>.83</td>
</tr>
<tr>
<td>Hypertension, No. (%)</td>
<td>672 (50)</td>
<td>668 (52)</td>
<td>.46</td>
</tr>
<tr>
<td>Tobacco use, No. (%)</td>
<td>181 (12)</td>
<td>152 (12)</td>
<td>.95</td>
</tr>
<tr>
<td><strong>Cardiac variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior coronary artery disease, No. (%)</td>
<td>602 (45)</td>
<td>659 (49)</td>
<td>.79</td>
</tr>
<tr>
<td>Prior coronary artery bypass graft, No. (%)</td>
<td>281 (19)</td>
<td>235 (17)</td>
<td>.42</td>
</tr>
<tr>
<td>Prior percutaneous coronary intervention, No. (%)</td>
<td>166 (12)</td>
<td>147 (11)</td>
<td>.35</td>
</tr>
<tr>
<td>Prior Q-wave MI, No. (%)</td>
<td>194 (14)</td>
<td>206 (15)</td>
<td>.52</td>
</tr>
<tr>
<td>Atrial fibrillation, No. (%)</td>
<td>21 (2)</td>
<td>24 (2)</td>
<td>.95</td>
</tr>
<tr>
<td>Congestive heart failure, No. (%)</td>
<td>79 (5)</td>
<td>89 (7)</td>
<td>.43</td>
</tr>
<tr>
<td><strong>Medication use</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digoxin use, No. (%)</td>
<td>115 (9)</td>
<td>114 (9)</td>
<td>.94</td>
</tr>
<tr>
<td>β-Blocker use, No. (%)</td>
<td>262 (20)</td>
<td>259 (20)</td>
<td>.79</td>
</tr>
<tr>
<td>Diltiazem/verapamil use, No. (%)</td>
<td>223 (17)</td>
<td>223 (17)</td>
<td>&gt;.09</td>
</tr>
<tr>
<td>Nifedipine use, No. (%)</td>
<td>127 (9)</td>
<td>144 (11)</td>
<td>.26</td>
</tr>
<tr>
<td>Lipid-lowering therapy, No. (%)</td>
<td>281 (21)</td>
<td>271 (20)</td>
<td>.63</td>
</tr>
<tr>
<td>ACE inhibitor use, No. (%)</td>
<td>200 (15)</td>
<td>214 (16)</td>
<td>.79</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>malignant died</th>
<th>malignant survived</th>
<th>benign died</th>
<th>benign survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>35</td>
<td>59</td>
<td>47</td>
<td>112</td>
</tr>
<tr>
<td>Glamorgan</td>
<td>42</td>
<td>77</td>
<td>26</td>
<td>76</td>
</tr>
</tbody>
</table>

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

→ We are mostly interested in the interactions!

Log-linear models

The saturated model:

<table>
<thead>
<tr>
<th>variable</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.00</td>
</tr>
<tr>
<td>center</td>
<td>0.42</td>
</tr>
<tr>
<td>grade</td>
<td>0.18</td>
</tr>
<tr>
<td>survival</td>
<td>0.01</td>
</tr>
<tr>
<td>center × grade</td>
<td>0.02</td>
</tr>
<tr>
<td>center × survival</td>
<td>0.76</td>
</tr>
<tr>
<td>grade × survival</td>
<td>0.20</td>
</tr>
<tr>
<td>grade × center × survival</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Log-linear models

A sub-model:

<table>
<thead>
<tr>
<th>variable</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.00</td>
</tr>
<tr>
<td>center</td>
<td>0.08</td>
</tr>
<tr>
<td>grade</td>
<td>0.15</td>
</tr>
<tr>
<td>survival</td>
<td>0.00</td>
</tr>
<tr>
<td>center × grade</td>
<td>0.00</td>
</tr>
<tr>
<td>grade × survival</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Survival Analysis
Survival analysis

Survival analysis: Study of durations between events

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

![Diagram of survival analysis with time milestones](image)

**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, . . .

![Diagram of censoring problem with time milestones](image)
Basic goals of survival analysis

1. To estimate and interpret survival characteristics
   → Kaplan-Meier plots

2. To compare survival in different groups
   → Log-rank test

3. To assess the relationship of explanatory variables to survival
   → Cox regression model

Survival function

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

→ \( S(t) \) 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 $t_j$.

If there are no censored observations, this reduces to

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

Example

![Graph showing Kaplan-Meier estimates for Control and 6-MP groups with time of remission in weeks on the x-axis and \( \hat{S}(t) \) on the y-axis. The blue line represents the Control group, and the red line represents the 6-MP group. The graph shows multiple steps, indicating censored and observed events.]
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 \( \chi^2 \)-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

\[
\logRT = \frac{Q^2}{\text{Var}(Q)} \sim \chi_1^2
\]
### Example

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>treat=6-MP</td>
<td>21</td>
<td>9</td>
<td>19.3</td>
<td>5.46</td>
<td>16.8</td>
</tr>
<tr>
<td>treat=control</td>
<td>21</td>
<td>21</td>
<td>10.7</td>
<td>9.77</td>
<td>16.8</td>
</tr>
</tbody>
</table>

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

\[
\logRT \sim \chi^2_{k-1}
\]

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

![Exponential and Weibull hazard functions](image-url)
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 \( \beta \) is a regression coefficient.

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
\[ RR = \frac{h(t|x = 1)}{h(t|x = 0)} = e^{\beta} \]
Cox regression model

→ Interpretation of the parameters:

\[ \beta > 0 \quad \text{RR} > 1 \quad \text{and} \quad h(t|x = 1) > h(t|x = 0) \]
\[ \beta = 0 \quad \text{RR} = 1 \quad \text{and} \quad h(t|x = 1) = h(t|x = 0) \]
\[ \beta < 0 \quad \text{RR} < 1 \quad \text{and} \quad h(t|x = 1) < h(t|x = 0) \]

→ Hypothesis of interest:

\[ H_0 : \beta = 0 \text{ (no treatment effect)} \]
\[ H_a : \beta \neq 0 \text{ (treatment influences survival)} \]

Example

\[
\begin{array}{cccccc}
\text{coef} & \exp(\text{coef}) & \text{se(\text{coef})} & z & p \\
treatcontrol & 1.57 & 4.82 & 0.412 & 3.81 & 0.00014 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\exp(\text{coef}) & \exp(-\text{\text{coef})} & \text{lower .95} & \text{upper .95} \\
treatcontrol & 4.82 & 0.208 & 2.15 & 10.8 \\
\end{array}
\]
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).

<table>
<thead>
<tr>
<th></th>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>agpresent</td>
<td>-1.069</td>
<td>0.343</td>
<td>0.429</td>
<td>-2.49</td>
<td>0.0130</td>
</tr>
<tr>
<td>log(wbc)</td>
<td>0.368</td>
<td>1.444</td>
<td>0.136</td>
<td>2.70</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>exp(coef)</th>
<th>exp(-coef)</th>
<th>lower .95</th>
<th>upper .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>agpresent</td>
<td>0.343</td>
<td>2.913</td>
<td>0.148</td>
<td>0.796</td>
</tr>
<tr>
<td>log(wbc)</td>
<td>1.444</td>
<td>0.692</td>
<td>1.106</td>
<td>1.886</td>
</tr>
</tbody>
</table>

Classification and Regression Trees
Example 1

[Scatter plot with markers labeled Outcome A, B, and C]
Example 1

Marker 1

Marker 2

Outcome A

Outcome B

Outcome C

Marker 1 < 0.065

Marker 2 < 10.54

Outcome C

151/0/0

Outcome B

0/98/0

Outcome A

0/0/323
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.
Again, suppose that we have a scalar outcome, $Y$, and a $p$-vector of explanatory variables, $X$. Now assume $Y \in \mathbb{R}$.

A regression tree partitions the $X$-space into disjoint regions $A_k$ and provides a fitted value $E(Y|X \in A_k)$ 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:
   $\rightarrow n - 1$ possible splits.

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

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

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

Trees

![Diagram of a tree structure with relative error on the y-axis and size on the x-axis.](image)
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).

Example: Low Birth Weight Data

<table>
<thead>
<tr>
<th>lwt &lt; 109.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>uah</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>n=42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2405</th>
</tr>
</thead>
<tbody>
<tr>
<td>smoke = b</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>n=17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2960</th>
</tr>
</thead>
<tbody>
<tr>
<td>race = bc</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>n=47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3097</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>n=44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3531</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>n=39</td>
</tr>
</tbody>
</table>
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
• ...