Biostatistics week 1BIO49

Karl Oskar Ekvall

Fall 2021

Karolinska Institutet

Table of contents

Table of contents

1. Simple linear regression

Experiment with fixed dose

We want to investigate the effect of dose on response.

We consider doses

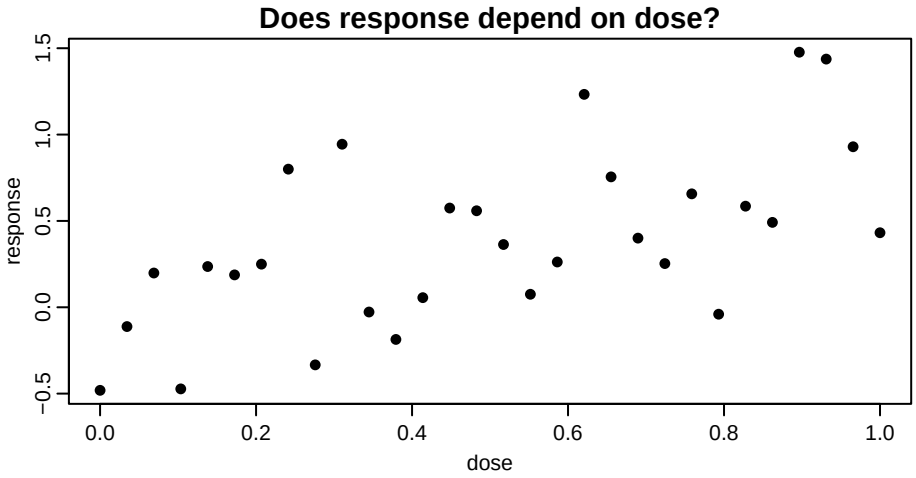
```
(x <- seq(0, 1, length.out = 30))
```

```
##  [1] 0.00000000 0.03448276 0.06896552 0.10344828 0.13793103 0.17241379
##  [7] 0.20689655 0.24137931 0.27586207 0.31034483 0.34482759 0.37931034
## [13] 0.41379310 0.44827586 0.48275862 0.51724138 0.55172414 0.58620690
## [19] 0.62068966 0.65517241 0.68965517 0.72413793 0.75862069 0.79310345
## [25] 0.82758621 0.86206897 0.89655172 0.93103448 0.96551724 1.00000000
```

For each dose x_i, we select one patient at random from the population of interest.

For each patient we observe the response y_i, $i = 1, \ldots, 30$.

Dose and response



A model for the effect of dose on response

Before we perform our experiment, the response of the ith person is a random variable Y_i whose distribution (potentially) depends on x_i.

The particular value y_i we observe depends on who were selected in the random sampling; not every person responds the same, there may be some measurement error, etc.

A *simple linear regression* model assumes

$$E(Y_i) = \mu(x_i) = \alpha + \beta x_i.$$

- μ is a function, and $\mu(x_i)$ is the value of that function evaluated at x_i
- α and β are unknown constants (parameters).

Interpreting the parameters

The parameters are interpreted using the equation

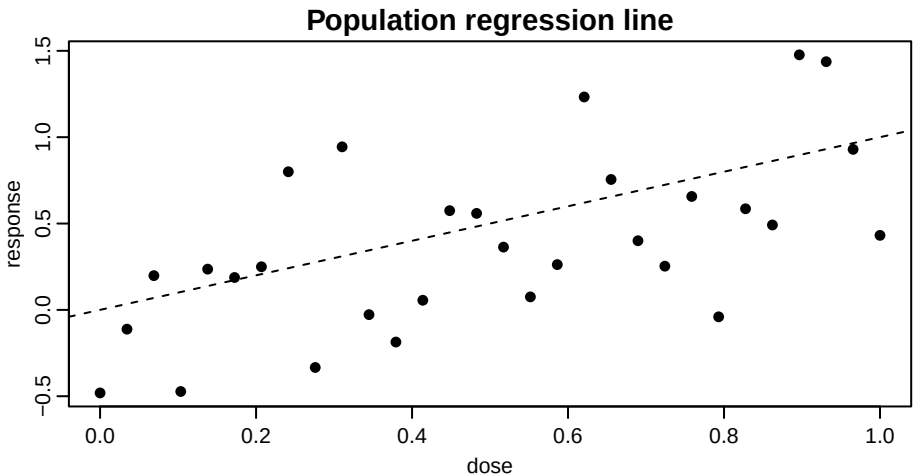
$$E(Y_i) = \alpha + \beta x_i.$$

- If x_i increases by one unit, Y_i increases by β units on average
- If $x_i = 0$, the mean of Y_i is α
- In the dose and response setting, β is the treatment effect and α is the average response for an untreated person

The population regression line

In practice the true regression line (dashed) is unknown because α and β are.

We only know it here because I generated the data in R (it's not "real" data).



In *ordinary least squares* (OLS) the parameters α and β are estimated by the a and b which minimize the *sum of squared residuals*

$$RSS = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

We often denote those a and b by $\hat{\alpha}$ and $\hat{\beta}$, respectively.

The estimated regression line is

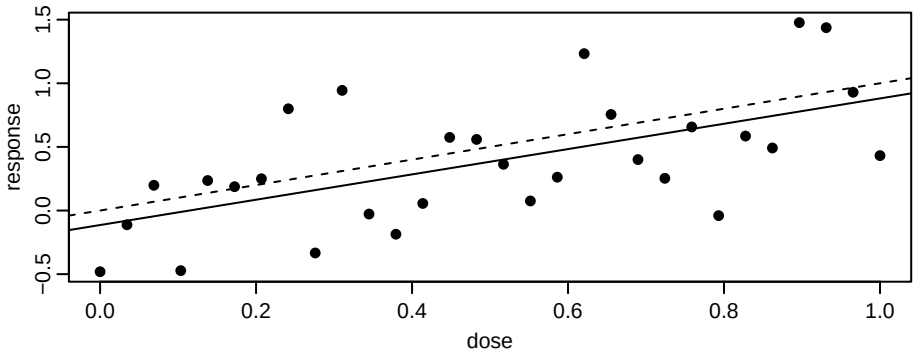
$$\hat{\mu}(x_i) = \hat{\alpha} + \hat{\beta} x_i.$$

Least squares estimates in R

```
fit <- lm(y ~ x); coef(fit)
```

```
## (Intercept)           x
##  -0.1140490   0.9948813
```

```
plot(y ~ x, ylab = "response", xlab = "dose"); abline(fit); abline(a = 0, b = 1, lty = 2)
```



The * indicates a slide with material that will not be tested

To find the least squares estimates, we minimize the Objective function:

$$SSR(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

First order optimality conditions:

$$\frac{\partial SSR(a, b)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial SSR(a, b)}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = -2 \sum_{i=1} (y_i - a) x_i + 2b \sum_{i=1}^{n} x_i^2 = 0$$

Solving the first order conditions gives

$$\hat{\alpha} = a = \bar{y} - b\bar{x},$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} x_i$ and $\bar{x}$ are sample averages.

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2},$$

where $s_{xy} = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ and s_x^2 are the sample covariance and variance, respectively.

Example with binary predictor

Suppose x_i is 1 if the ith patient receives treatment A and 0 if they receive treatment B.

Then for patients receiving treatment A

$$E(Y_i) = \alpha,$$

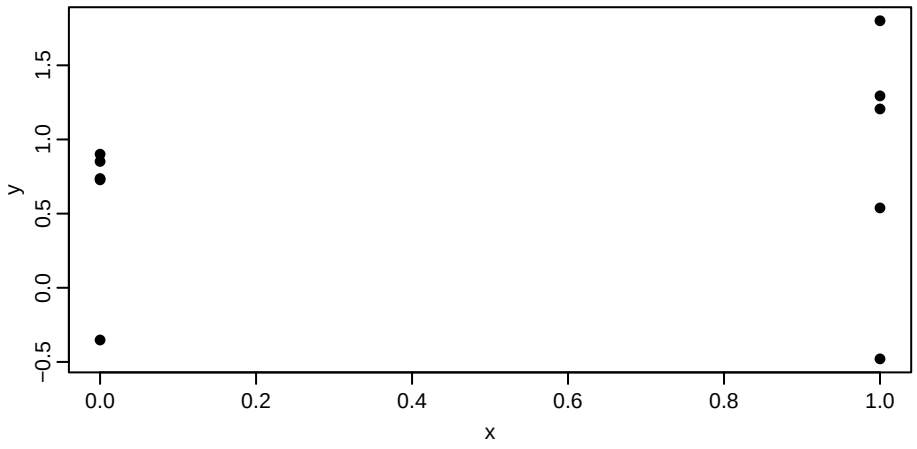
and for patients receiving treatment B

$$E(Y_i) = \alpha + \beta.$$

If $\beta = 0$, the mean of Y_i is the same in both groups.

Example with binary predictor

```
x <- c(0, 0, 0, 0, 0, 1, 1, 1, 1, 1); y <- rnorm(10, 0.5 * x) # alpha = 0, beta = 0.5
plot(x, y)
```



Example with binary predictor

```
b <- cov(x, y) / var(x); a <- mean(y) - b * mean(x)
c(a, b) # Parameter estimates
```

```
## [1] 0.5728966 0.2988240
```

```
c(a, a + b) # Group mean estimates
```

```
## [1] 0.5728966 0.8717206
```

```
c(mean(y[x == 0]), mean(y[x == 1])) # Sample group means
```

```
## [1] 0.5728966 0.8717206
```

Example with binary predictor

Estimating means in two different groups is a special case of linear regression!

In fact, we will see that t-tests, ANOVA, and confidence intervals for the mean are obtained as special cases of inference in linear regression.

- Some work to do before we get there

We have point estimates $\hat{\alpha}$ and $\hat{\beta}$ and now want to quantify our uncertainty:

- Are the estimates reliable?
- What are the standard errors of the estimates?
- How to create confidence intervals for α and β?
- Are α and β significantly different from zero?

Define the random error term U_i, $i = 1, \ldots, n$, by

$$U_i = Y_i - \mu(x_i),$$

or equivalently

$$Y_i = \alpha + \beta x_i + U_i.$$

The error term U_i is an unobservable random variable:

- We do not know its realized value $u_i = y_i - \mu(x_i)$ because α and β are unknown.
- We do observe the *residual* $r_i = y_i - \hat{\mu}(x_i) \neq u_i$.

To make inferences about α and β, we assume

1. The mean of U_i is zero; $E(U_i) = 0$
2. The variance of U_i is constant (homoscedastic, not depending on i or x_i); $\mathrm{var}(U_i) = \sigma^2 > 0$

Assumption 1 ensures the estimators are unbiased:

$$E(\hat{\alpha}) = \alpha; \;\; E(\hat{\beta}) = \beta$$

Together, assumption 1 and 2 ensure, by a central limit theorem,

$$\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{\sigma^2}{(n-1)s_x^2}\right); \quad \hat{\alpha} \overset{approx.}{\sim} N\left(\alpha, \frac{\sigma^2(s_x^2 + \bar{x}^2)}{(n-1)s_x^2}\right)$$

- The variance tends to zero, which implies the estimators are consistent (close to the true values with increasing probability as $n \to \infty$)
- We can use the approximate normality to calculate standard errors, tests, and confidence intervals

Inference with normaly distributed error term and known σ^2

If the error term is **normally distributed**, then $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed (no approximation).

If, moreover, the variance σ^2 of the error is **known**, a 95 % confidence interval is:

$$\hat{\beta} \pm 1.96 \times \sqrt{\mathrm{var}(\hat{\beta})} = \hat{\beta} \pm 1.96 \frac{\sigma}{\sqrt{(n-1)}s_x}.$$

Recall that $1.96 = \texttt{qnorm}(\texttt{0.975})$ is the 0.975th quantile of the standard normal distribution.

Estimating σ^2

If the error term is normally distributed with mean zero and *unknown* variance σ^2, we cannot use the confidence interval on the previous slide.

To get a confidence interval, we will first need to estimate σ^2.

We will use

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{2} r_i^2 = \frac{1}{n-2}SSR,$$

where as before $r_i = y_i - \hat{\mu}(x_i)$ is the residual.

Dividing by $n-2$ ensures $E(S_r^2) = \sigma^2$ so the estimator is unbiased.

The standard error of $\hat{\beta}$

$$\text{se}(\hat{\beta}) = \sqrt{\frac{s_r^2}{(n-1)s_x^2}}$$

is an estimate of

$$\sqrt{\text{var}(\hat{\beta})} = \sqrt{\frac{\sigma^2}{(n-1)s_x^2}}.$$

The statistic

$$\frac{\hat{\beta} - \beta}{S_r^2 / \{(n-1)s_x^2\}} \sim t_{n-2}.$$

and therefore a 95 % confidence interval is

$$\hat{\beta} \pm \texttt{qt}(0.975, n-2) \times \text{se}(\hat{\beta}).$$

Example with binary predictor

```
b <- cov(x, y) / var(x)
a <- mean(y) - b * mean(x)
res <- y - a - x * b
s2r <- sum(res^2) / (10 - 2)
se_b <- sqrt(s2r / ((10 - 1) * var(x)))
b + c(-1, 1) * qt(0.975, 10 - 2) * se_b
```

```
## [1] -0.7553887  1.3530367
```

```
confint(lm(y ~ x))
```

```
##                  2.5 %   97.5 %
## (Intercept) -0.1725444 1.318338
## x           -0.7553887 1.353037
```

We can use the fact

$$\frac{\hat{\beta} - \beta}{S_r^2/\{(n-1)s_x^2\}} \sim t_{n-2}.$$

to test the null hypothesis $\beta = \beta_0$ for any β_0 of interest.

Suppose we want to test the null hypothesis $\beta = 0$. Under the null hypothesis,

$$T = \frac{\hat{\beta}}{S_r^2/\{(n-1)s_x^2\}} \sim t_{n-2}$$

Recall the intuition behind hypothesis testing:

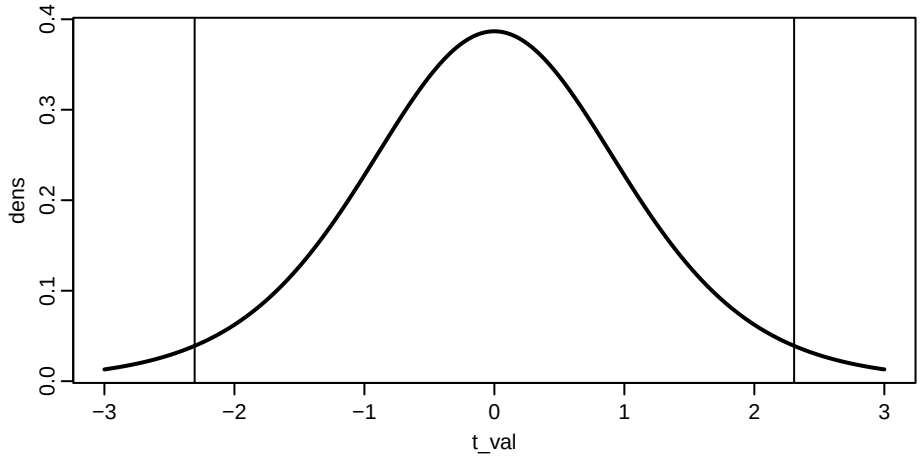
Reject if what we observe is unlikely under the null hypothesis

More formally, we reject on the 5 % level if under the null hypothesis

$$P(|T| \geq |t|) \leq 0.05,$$

where $|t|$ is the value of T observed in our sample. That is, we reject if

$$t > \texttt{qt}(0.975, n-2) \text{ or } t < \texttt{qt}(0.025, n-2) \iff |t| \geq \texttt{qt}(0.975, n-2)$$




```
c(qt(0.025, 8), qt(0.975, 8))
```

```
## [1] -2.306004  2.306004
```

Example with binary predictor

```r
# Test beta = 0
t <- b / se_b
abs(t)
```

```
## [1] 0.6536531
```

```r
qt(0.975, 10 - 2)
```

```
## [1] 2.306004
```

```r
2 * pt(abs(t), 10 - 2, lower = F) # p-value
```

```
## [1] 0.5316699
```

Example with binary predictor

```
summary(lm(y ~ x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3510 -0.2114  0.2212  0.3323  0.9286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5729     0.3233   1.772    0.114
## x             0.2988     0.4572   0.654    0.532
##
## Residual standard error: 0.7228 on 8 degrees of freedom
## Multiple R-squared:  0.0507, Adjusted R-squared:  -0.06796
## F-statistic: 0.4273 on 1 and 8 DF,  p-value: 0.5317
```

Example with binary predictor

Recall, $\beta = 0$ is the same as the two groups having the same mean.

```
t.test(y[x == 1], y[x == 0], var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  y[x == 1] and y[x == 0]
## t = 0.65365, df = 8, p-value = 0.5317
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7553887  1.3530367
## sample estimates:
## mean of x mean of y
## 0.8717206 0.5728966
```

2. Simple logistic regression

Suppose Y_i is binary (Bernoulli); that is, it takes the value 1 with probability p_i and the value 0 with probability $1 - p_i$.

Then the linear regression model

$$p_i = E(Y_i) = \mu(x_i) = \alpha + \beta x_i$$

is also known as the *linear probability model*.

This model can be useful, but it has an important drawback.

In many settings, there is no upper bound on the possible values of x_i.

For any $\beta \neq 0$, large enough $|x_i|$ can lead to $p_i = \alpha + \beta x_i > 1$ or $p_i < 0$, which are impossible!

A common solution is to assume instead

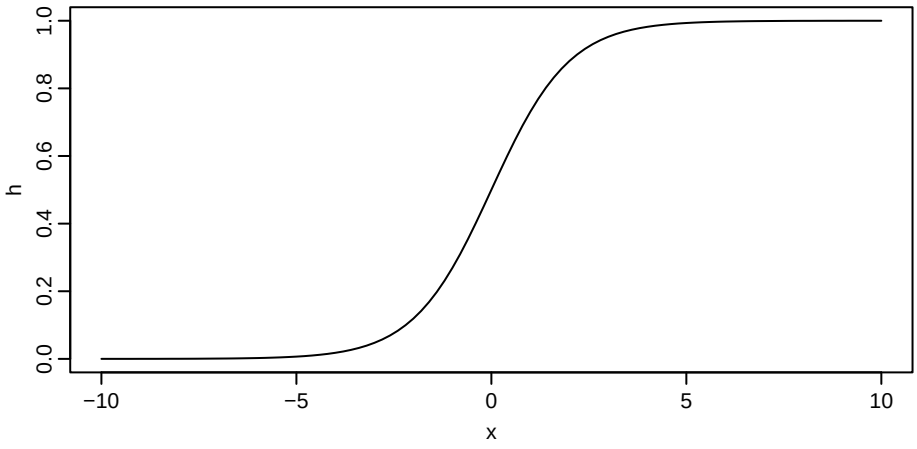
$$p_i = E(Y_i) = \mu(x_i) = \frac{1}{1 - \exp(-\alpha - \beta x_i)} = h(\alpha + \beta x_i),$$

where h is known as the logistic function, defined by

$$h(z) = \frac{1}{1 + \exp(-z)}.$$

The logistic function

```
h <- function(z){1 / (1 + exp(-z))}
plot(h, -10, 10)
```



No matter what $\alpha + \beta x_i$ is, $\mu(x_i) = h(\alpha + \beta x_i)$ is a value between 0 and 1, as required by the Bernoulli distribution.

The *marginal effect* of x_i is

$$\frac{\partial \mu(x_i)}{\partial x_i} = \beta h'(\alpha + \beta x_i) = \beta \frac{\exp(-\alpha - \beta x_i)}{\{1 + \exp(-\alpha - \beta x_i)\}^2}.$$

This is not easy to interpret, but:

- It is zero if $\beta = 0$
- It has the same sign as β
- It is smaller for large $|\alpha + \beta x_i|$

Interpretation with binary predictor

If x_i is binary, then

$$\frac{P(Y_i = 1 \mid x_i = 1)/P(Y_i = 0 \mid x_i = 1)}{P(Y_i = 1 \mid x_i = 0)/P(Y_i = 0 \mid x_i = 0)} = e^{\beta},$$

so e^{β} is an odds ratio.

For example, suppose:

- $Y_i = 1$ if patient i recovers and $Y_i = 0$ otherwise
- $x_i = 1$ if patient i receives treatment and 0 otherwise
- $\beta = 0.2$

Then the odds of surviving is $\exp(0.2) \approx 1.2$ times higher if the patient is treated.

It is possible to estimate α and β by least squares; that is, by the a and b which minimize

$$\sum_{i=1}^{n}\{y_i - \mu(x_i)\}^2 = \sum_{i=1}^{n}\{y_i - h(a + bx_i)\}^2.$$

Let us look at an example in R.

Logistic regression by least squares*

```r
x <- runif(100, -1, 1)
y <- rbinom(100, 1, prob = h(x))
summary(nls(y ~ 1 / (1 + exp(-a - b * x)), start = list(a = 0, b = 1)))
```

```
##
## Formula: y ~ 1/(1 + exp(-a - b * x))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a -0.08675    0.21101  -0.411  0.68190
## b  1.11371    0.41198   2.703  0.00809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4835 on 98 degrees of freedom
##
## Number of iterations to convergence: 3
## Achieved convergence tolerance: 3.385e-06
```

Logistic regression by maximum likelihood

Least squares is not the most common way to fit a logistic regression model.

Instead, the standard fitting procedure is *maximum likelihood*.

Maximum likelihood estimates are the parameter values which maximize the probability of observing the data we did observe.

Our observed responses are $y_1, \ldots, y_n$. The probability of observing those values before we performed the experiment sampling was

$$P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\} \cap \cdots \cap \{Y_n = y_n\})$$

Because the events are independent (random sampling), this is equal to

$$P(Y_1 = y_1)P(Y_2 = y_2) \cdots P(Y_n = y_n).$$

The probability $P(Y_i = y_i)$ is the mass function of Y_i evaluated at y_i, which is

$$f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i},$$

where $p_i = \mu(x_i) = h(\alpha + \beta x_i)$.

Therefore,

$$P(Y_1 = y_1)P(Y_2 = y_2) \cdots P(Y_n = y_n) = \prod_{i=1}^{n} h(\alpha + \beta x_i)^{y_i}\{1 - h(\alpha + \beta x_i)\}^{1-y_i}.$$

The function L defined by

$$L(\alpha, \beta) = \prod_{i=1}^{n} h(\alpha + \beta x_i)^{y_i} \{1 - h(\alpha + \beta x_i)\}^{1-y_i} \quad (\star)$$

is called the likelihood function. Observe, here α and β are arguments to the function, not fixed at the true values.

- The α and β which maximize L are the maximum likelihood estimates of the true α and β.
- The maximum likelihood estimates depend on the data because L does.
- The maximizers can be computed when given data, so they are statistics.
- Before sampling when the data are random, the maximizers of L are also random

Example in R

```
summary(glm(y ~ x, family = binomial))
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5864  -1.0431  -0.7659   1.0714   1.6619
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.08065    0.20886  -0.386  0.69939
## x            1.09520    0.39387   2.781  0.00543 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.47  on 99  degrees of freedom
```

What you need to know about maximum likelihood

You should

- Know what maximum likelihood estimation is (with discrete variables)
- How to do it in R for the models we discuss
- How to interpret the output from R (estimate, se, z-value, p-value)

How to find the estimates, derive the formulas for standard errors, etc. are topics for more advanced statistics classes.

3. The role of covariates

Covariates

A covariate is a variable (potentially) related to the response.

Suppose we randomly sample n people in Stockholm and record

- $Y_i = 1$ if person i had COVID-19 in the last 30 days, zero otherwise
- $X_{1i} = 1$ if person i were fully vaccinated 30 days ago, zero otherwise
- $X_{2i} =$ the age of person i

We are interested in the effect of vaccination on the probability of COVID-19.

- Which covariates should we include in the analysis, and how?

Different roles of covariates

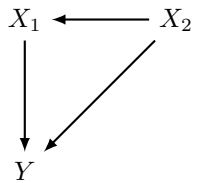
It is common to call the effect we are interested in the treatment effect.

In the example, the effect of vaccination on probability of COVID-19 is the treatment effect.

Age may also be important.

Causal inference – distinguish causal effects from spurious correlations

A DAG (directed acyclic graph) can help clarify:



- Nodes and edges
- Arrows indicate direction of causality
  - Parents and children, descendants and ancestors
- No particular model is assumed

Paths

A *direct path* indicates a causal relationship

- $X_1 \longrightarrow Y$ and $X_2 \longrightarrow X_1 \longrightarrow Y$ (a *chain*)
  - X_1 has a direct effect and is a *mediator* for the effect of X_2

A *backdoor path* from treatment to response can lead to spurious correlation

- $X_1 \longleftarrow X_2 \longrightarrow Y$ (a *fork*)
  - X_2 is called a *confounder*
  - If vaccine has no effect on COVID-19 and age increases probability of vaccine and probability of COVID-19, there will be a spurious negative correlation between vaccine and COVID-19.

Blocking a path

It is clear we have to account for age, but how?

We can block the backdoor path by *conditioning* on age

- consider the effect of the vaccine for fixed age

Let's look at an example in R.

Example in R

```r
n <- 1e4
age <- sample(20:90, n, replace = T)
old <- as.numeric(age >= 60)
vaccine <- rbinom(n, 1, h(2 * old))
covid <- rbinom(n, 1, h(-2 - vaccine + 3 * old))

# True probabilities
c("y,v" = h(-2 - 1), "y,u" = h(-2), "o,v" = h(-2 - 1 + 3), "o, u" = h(-2 + 3))
```

```
##        y,v        y,u        o,v       o, u
## 0.04742587 0.11920292 0.50000000 0.73105858
```

Conditioning on a confounder

A naive comparison of means without conditioning on age indicates vaccine has negative effect:

```
mean(covid[vaccine == 1]) -  mean(covid[vaccine == 0])
```

```
## [1] 0.08403924
```

Conditioning on age:

```
# For old people
mean(covid[vaccine == 1 & old == 1]) - mean(covid[vaccine == 0 & old == 1])
```

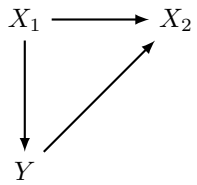
```
## [1] -0.246411
```

```
# For young people
mean(covid[vaccine == 1 & old == 0]) - mean(covid[vaccine == 0 & old == 0])
```

```
## [1] -0.08602107
```

Another type of covariate

Consider the DAG:



- X_1 is vaccination, X_2 is travel (no effect on Y!), and Y is COVID-19

There is a backdor path from X_1 to Y, and X_2 is a *collider* on that path.

When a path has a collider on it, that path is closed and you should not condition on the collider.

Example in R

```
vaccine <- rbinom(n, 1, 0.5)
covid <- rbinom(n, 1, h(-2 * vaccine))
travel <- rbinom(n, 1, h(5 * vaccine - 10 * covid))
mean(covid[vaccine == 1]) - mean(covid[vaccine == 0])
```

```
## [1] -0.377719
```

```
mean(covid[vaccine == 1 & travel == 1]) - mean(covid[vaccine == 0 & travel == 1])
```

```
## [1] 0.0009378664
```

```
mean(covid[vaccine == 1 & travel == 0]) - mean(covid[vaccine == 0 & travel == 0])
```

```
## [1] 0.2901262
```

Backdoor paths in general

You want to close all backdoor paths from treatment to response.

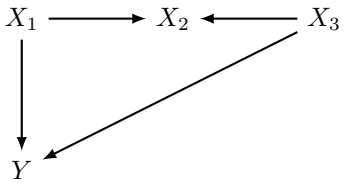
May need to condition on several variables, let's call the collection of variable we condition on Z:

A path is blocked if:

- it contains a chain or fork whose middle node is in Z, or
- it contains a collider such that neither the middle node nor or any of its descendatnts are in Z

It may not be possible to block all backdoor paths!

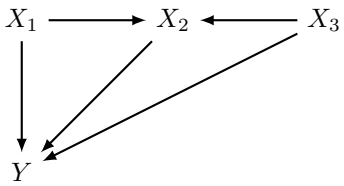
Example with three covariates



The path $X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y$ contains one fork and one collider

- It is closed if we do not condition because it contains a collider
- It is closed if we condition on X_3 only
- It is open if we condition on X_2 only
- It is closed if we condition on X_2 and X_3

More difficult example with three covariates



(1) The path $X_1 \to X_2 \leftarrow X_3 \to Y$ contains one fork and one collider.

(2) The path $X_1 \to X_2 \to Y$ contains a fork.

- Path (2) is open unless we condition on X_2
- Conditioning on X_2 opens path (1), so we close it by conditioning on X_3.

Things to address

To make this useful in practice we need to understand:

- How to examine the effect of a random, numeric covariate on a random response.
- How to condition on several variables at the same time.

We can address all of these with regression methods.

Random covariates can be used in regression, but the formal motivation is different.

We now say the conditional mean of Y_i given X_i is

$$E(Y_i \mid X_i) = \mu(X_i),$$

for some known function μ. In linear regression $\mu(X_i) = \alpha + \beta X_i$ and in logistic regression $\mu(X_i) = 1/\{1 + \exp(-\alpha - \beta X_i)\}$.

- Treatment of conditional distributions is outside our scope, suffices to know the same methods work.

We can interpret the conditional expectation

$$E(Y \mid X)$$

as the population average of Y for any given value of X. When we consider this expectation for a particular value x of X, it is common to write

$$E(Y \mid X = x).$$

Logistic regression with random covariate

```
(fit <- glm(covid ~ vaccine, family = binomial))
```

```
##
## Call:  glm(formula = covid ~ vaccine, family = binomial)
##
## Coefficients:
## (Intercept)      vaccine
##   -0.002737    -1.974659
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
## Null Deviance:       12460
## Residual Deviance: 10710     AIC: 10710
```

$$\hat{\mu}(\text{vacc}) = \hat{P}(\text{covid} \mid \text{vacc}) = \frac{1}{1 + \exp(0.00274 - 1.97 \times \text{vacc})}$$

Logistic regression with random covariate

```
a <- unname(coef(fit)[1]); b <- unname(coef(fit)[2])
h(a) # Estimate of P(covid | not vacc)
```

```
## [1] 0.4993157
```

```
h(a + b) # Estimate of P(covid | vacc)
```

```
## [1] 0.1215967
```

```
mean(covid[vaccine == 0])
```

```
## [1] 0.4993157
```

```
mean(covid[vaccine == 1])
```

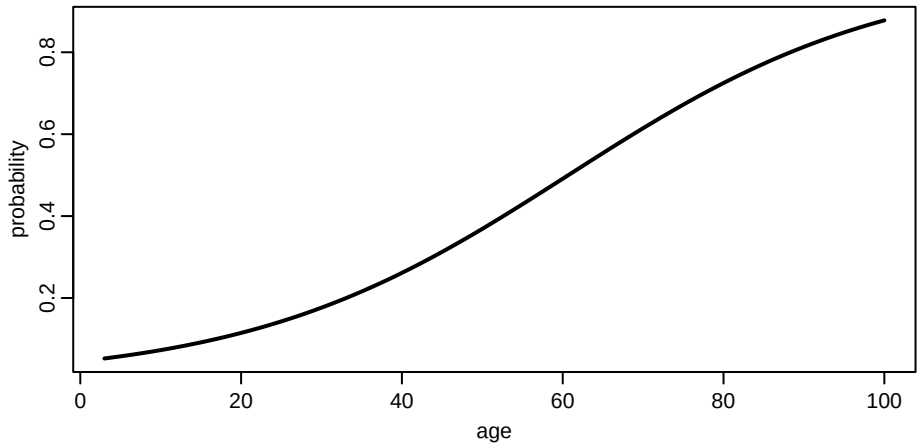
```
## [1] 0.1215967
```

Logistic regression with random numeric covariate

```r
severe <- rbinom(n, 1, h(-3 + 0.05 * age))
(fit <- glm(severe ~ age, family = binomial))
```

```
##
## Call:  glm(formula = severe ~ age, family = binomial)
##
## Coefficients:
## (Intercept)          age
##     -3.0470       0.0502
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
## Null Deviance:       13740
## Residual Deviance: 11590     AIC: 11590
```

$$\hat{\mu}(\text{age}) = \hat{P}(\text{severe} \mid \text{age}) = \frac{1}{1 + \exp(3.05 - 0.050 \times \text{age})}$$



4. Multiple regression

In *multiple regression*, we have a vector of p regressors $X_i = [X_{i1}, \ldots, X_{ip}]$ affecting the response.

Multiple linear regression assumes

$$E(Y \mid X_i) = \sum_{j=1}^{n} X_{ij}\beta_j = X_{i1}\beta_1 + \cdots + X_{ip}\beta_p.$$

Multiple logistic regression assumes

$$E(Y \mid X_i) = h\left(\sum_{j=1}^{n} X_{ij}\beta_j\right)$$

Parameter interpretation

It is common to let the first predictors $X_{i1} = 1$ for all i so that β_1 is an intercept (which we previously denoted α).

The parameter β_j indicates the effect of X_{ij} on the mean of Y_i *holding all the other regressors fixed*.

Including a random covariate as regressor is a way to condition on that covariate.

Example with binary variables

Consider the same example as before:

```
n <- 1e4
age <- sample(20:90, n, replace = T)
old <- as.numeric(age >= 60)
vaccine <- rbinom(n, 1, h(2 * old))
covid <- rbinom(n, 1, h(-2 - vaccine + 3 * old))
```

$$E(Y_i \mid X_i) = \frac{1}{1 + \exp(2 + \text{vacc} - 3 \times \text{old})}$$

We have $\beta = [\beta_1, \beta_2, \beta_3] = [-2, -1, 3]$.

Example with binary variables

If we do not include old as regressor:

```
glm(covid ~ vaccine, family = binomial)
```

```
##
## Call:  glm(formula = covid ~ vaccine, family = binomial)
##
## Coefficients:
## (Intercept)      vaccine
##     -1.3021       0.4402
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
## Null Deviance:       11660
## Residual Deviance: 11580     AIC: 11580
```

The estimate suffers from *omitted variable bias*.

Example with binary variables

Similar results if age is included instead of old.

If we do include old as regressor:

```
glm(covid ~ vaccine + old, family = binomial)
```

```
##
## Call:  glm(formula = covid ~ vaccine + old, family = binomial)
##
## Coefficients:
## (Intercept)      vaccine          old
##      -1.996       -1.005        2.940
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9997 Residual
## Null Deviance:       11660
## Residual Deviance: 9048  AIC: 9054
```

No omitted variable bias.

Example with binary response, numeric covariate

```
vaccine <- rbinom(n, 1, h(0.01 * age))
covid <- rbinom(n, 1, h(-2 - vaccine + 0.05 * age))
glm(covid ~ vaccine, family = binomial)
```

```
##
## Call:  glm(formula = covid ~ vaccine, family = binomial)
##
## Coefficients:
## (Intercept)      vaccine
##      0.4777      -0.6237
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
## Null Deviance:       13850
## Residual Deviance: 13630     AIC: 13630
```

Example with binary response, numeric covariate

```
glm(covid ~ vaccine + age, family = binomial)
```

```
##
## Call:  glm(formula = covid ~ vaccine + age, family = binomial)
##
## Coefficients:
## (Intercept)      vaccine          age
##    -1.95467     -1.01148      0.04902
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9997 Residual
## Null Deviance:       13850
## Residual Deviance: 11670     AIC: 11680
```

Modeling choices

We have seen a DAG can help decide which covariates should be conditioned on.

But how do we know in practice how they affect the response? That is, how do we know which model to pick?

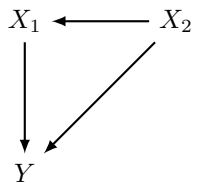
We don't! Many models are consistent with the same DAG.

Difference between DAG and model

For example, all of

$$\mu(X) = h(\beta_1 X_1 + \beta_2 X_2)$$
$$\mu(X) = h(\beta_1 X_1 + \beta_2 X_2^2)$$
$$\mu(X) = |\beta_1 X_1|/(1 + |\beta_1 X_1| + |\beta_2 X_2|)$$

are consistent with the following DAG



We need tools for *model selection*.

Let us first focus on testing the importance of some regressors assuming the rest of the model is correct.

For example, we want to compare

$$
\begin{aligned}
(M1) \quad & \mu(X) = \beta_1 X_1 \\
(M2) \quad & \mu(X) = \beta_1 X_1 + \beta_2 X_2 \\
(M3) \quad & \mu(X) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2
\end{aligned}
$$

All fit in our framework if we define $X_3 = X_1^2$, so no new fitting methods are needed.

Nested models

We say that models M1–M3 are *nested* since:

- M1 is a special case of M2 with $\beta_2 = 0$
- M2 is a special case of M3 with $\beta_3 = 0$
- M1 is a special case of M3 with $\beta_2 = \beta_3 = 0$

We can compare nested models using hypothesis tests.

Sum of squared residuals

We start with linear multiple regression where $\hat{\beta}$ is the $\beta = [\beta_1, \ldots, \beta_p]$ which minimizes

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2.$$

Define

$$SSR = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \hat{\beta}_j X_{ij} \right)^2 = \sum_{i=1}^{n} r_i^2.$$

Sum of squared residuals

Let SSR_1 be the sum of squared residuals for M1 and SSR_2 the sum of squared residuals from M2.

Why don't we just pick the model whose SSR is smaller?

Let $\tilde{\beta}$ be the estimate from M1 and $\hat{\beta}$ the estimate from M2.

$$SSR_1 = \sum_{i=1}^{n} \left(Y_i - \tilde{\beta}_1 X_{i1} - 0 X_{i2} \right)^2 \geq \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} \right)^2 = SSR_2.$$

Adding regressors (more flexibility) always gives lower SSR!

- Does not matter whether those regressors are actually related to the response

We can perform a hypothesis test instead.

Under the null hypothesis that the true $\beta_2 = 0$:

$$F = \frac{SSR_1 - SSR_2}{SSR_2/(n-2)} \sim \mathrm{F}_{1,n-2}.$$

Reject the null hypothesis if

$$F > \texttt{qf}(1-\alpha, 1, n-2),$$

where $\alpha \in (0,1)$ is the significance level and `qf` is the quantile function for the F-distribution.

More generally, let SSR_U be the sum of squared residuals for an *unrestricted* model and SSR_R the sum of squared resilduals for a *restricted* model which assumes some of the coefficients in the unrestricted model are zero. Then

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-p)} \sim \mathrm{F}_{q,n-p},$$

where q is the number of restrictions and p the number of regressors in the unrestricted model.

Reject the null hypothesis the coefficients are zero if

$$F > \mathtt{qf}(1-\alpha, q, n-p).$$

Example in R

```
n <- 45
x1 <- runif(n); x2 <- rexp(n)
y <- rnorm(n, mean = -0.1 + x1 + 0.1 * x2, sd = 0.5)
# Test both slope coefficients are zero
fit_UR <- lm(y ~ x1 + x2); fit_R <- lm(y ~ 1)
SSR_UR <- sum(residuals(fit_UR)^2); SSR_R <- sum(residuals(fit_R)^2)
((SSR_R - SSR_UR) / 2) / (SSR_UR / (n - 3))
```

```
## [1] 9.021929
```

```
qf(0.95, 2, n - 3)
```

```
## [1] 3.219942
```

Example in R

```
summary(fit_UR)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1185 -0.2297  0.0459  0.2775  1.1365
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.11928    0.18516  -0.644 0.522939
## x1           1.12097    0.26403   4.246 0.000118 ***
## x2           0.06993    0.09455   0.740 0.463636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.516 on 42 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.2672
## F-statistic: 9.022 on 2 and 42 DF,  p-value: 0.00055
```

Example in R

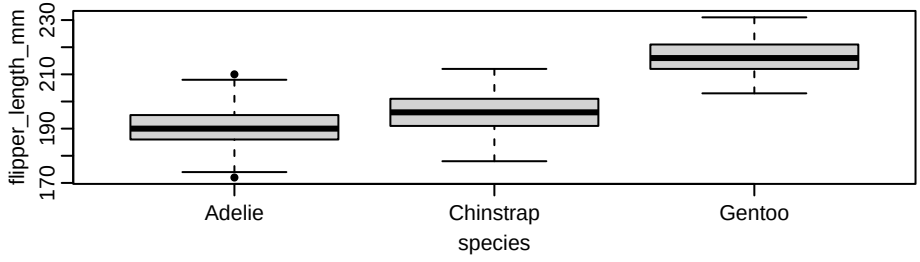
```
anova(fit_R, fit_UR)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x1 + x2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     44 15.989
## 2     42 11.184  2    4.8049 9.0219 0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Recall ANOVA can be used to test whether the means $\mu_1 = \cdots = \mu_p$ for p normally distributed populations / groups.

For example, are the flipper lengths the same for all three species?

```
library(palmerpenguins)
boxplot(flipper_length_mm ~ species, data = penguins)
```



Let $x_{i1} = 1$, $x_{i2} = 1$ if penguin i is Chinstrap and zero otherwise, and $x_{i3} = 1$ if penguin i is Gentoo and zero otherwise.

$$E(Y_i \mid X_i = x_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

The mean for Adelie penguins is $\mu_1 = \beta_1$, the mean for Chinstrap penguins is $\mu_2 = \beta_1 + \beta_2$, and the mean for Gentoo penguins is $\mu_3 = \beta_1 + \beta_3$.

The null hypothesis $\mu_1 = \mu_2 = \mu_3$ is the same as $\beta_2 = \beta_3 = 0$.

ANOVA as a special case

```r
y <- penguins$flipper_length_mm
x1 <- as.numeric(penguins$species == "Chinstrap")
x2 <- as.numeric(penguins$species == "Gentoo")
anova(lm(y ~ 1), lm(y ~ x1 + x2))
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x1 + x2
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1    341 67427
## 2    339 14953  2     52473 594.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA as a special case

```
anova(lm(flipper_length_mm ~ species, data = penguins))
```

```
## Analysis of Variance Table
##
## Response: flipper_length_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2  52473 26236.6   594.8 < 2.2e-16 ***
## Residuals 339  14953    44.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For logistic regression there are no residuals in the usual sense so other methods are needed.

We will use likelihood-based methods.

In multiple logistic regression, $\hat{\beta}$ is the β maximizing

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{y_i - 1}, \quad p_i = \mu(x_i) = h\left(\sum_{j=1}^{p} \beta_j x_{ij}\right).$$

Let $\tilde{\beta}$ be the estimate from a model which restricts some coefficients to zero. Then

$$L(\hat{\beta}) \geq L(\tilde{\beta}).$$

Adding regressors always leads to greater likelihood!

Under the null hypothesis, approximately for large n,

$$LLR = 2\log\left\{\frac{L(\hat{\beta})}{L(\tilde{\beta})}\right\} \sim \chi^2_q,$$

where q is the number of restrictions. So reject if

$$LLR > \texttt{qchisq}(1-\alpha, q).$$

Example in R

```
anova(glm(covid ~ vaccine, family = binomial),
      glm(covid ~ vaccine + age, family = binomial), test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: covid ~ vaccine
## Model 2: covid ~ vaccine + age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9998      13626
## 2      9997      11671  1   1954.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Non-nested models

What if we want to compare non-nested models?

- Hypothesis testing typically not useful
- Selecting the model with the greatest likelihood will generally lead to too large models

We can use *infomation criteria*.

Information criteria

Information criteria are essentially the likelihood with a penalty for the number of parameters:

Akaike's Information Criterion (AIC):

$$-2\log L(\hat{\beta}) + 2p.$$

Schwarz's Bayesian Criterion (BIC):

$$-2\log L(\hat{\beta}) + \log(n)p$$

- Pick the model with the smallest IC.

AIC or BIC

BIC will favor smaller models than AIC.

If the true model is among the candidates, BIC is likely to select it if the sample is large.

If the true model is not among the candidates, AIC is likely to select the model that comes closest if the sample is large.

Some people prefer AIC for prediction and BIC for inference, but there are no hard rules.

Example in R

```
AIC(lm(covid ~ vaccine + vaccine:age))
```

```
## [1] 13023.44
```

```
AIC(lm(covid ~ age + vaccine + I(age^2)))
```

```
## [1] 12272.72
```

Example in R

```
BIC(lm(covid ~ vaccine + vaccine:age))
```

```
## [1] 13052.28
```

```
BIC(lm(covid ~ age + vaccine + I(age^2)))
```

```
## [1] 12308.77
```

The term `age : vaccine` is called an interaction.

One can think of age as modifying the effect of vaccination.

$$E(Y \mid X) = \beta_1 + \beta_2 \texttt{vaccine} + \beta_3 \texttt{vaccine} \times \texttt{age}.$$

If β_3 is positive, then the effect of vaccination on the response increases with age.

The term $\texttt{I}(\texttt{age}^2)$ says

$$E(Y \mid X) = \beta_1 + \beta_2 \texttt{age} + \beta_3 \texttt{vaccine} + \beta_4 \texttt{age}^2.$$

If $\beta_4 < 0$, then the effect of age on the response decreases with age.

5. Survival analysis

Survival function

Survival analysis is concerned with the time T a randomly selected patient (or something or someone else of interest) survives.

Often depends on covariates.

The *survival function* is the function S defined by

$$S(t) = P(T > t).$$

Also called complementary distribution function since the complement of $T > t$ is $T \leq t$ and

$$S(t) = 1 - P(T \leq t) = 1 - F(t).$$

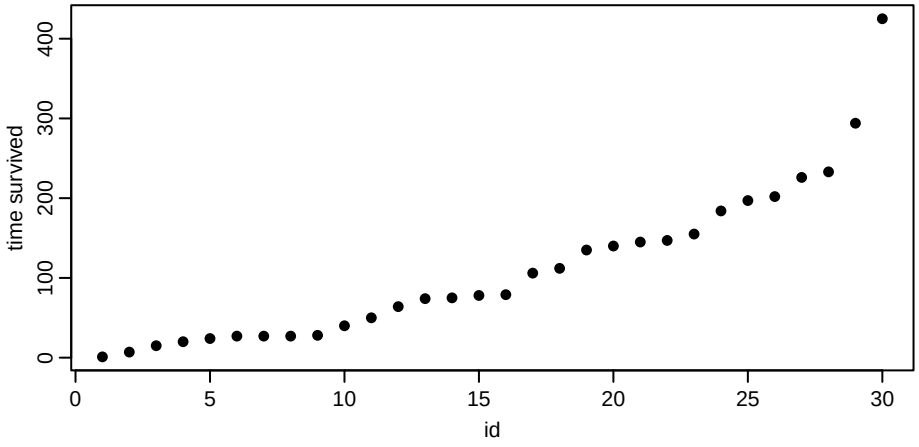
The most common estimator of $S(t)$ is the Kaplan–Meier estimator

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{\#\{\text{died at time } t_i\}}{\#\{\text{survived at least until } t_i\}}\right) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

It is a non-parametric estimate (does not assume a particular distribution).

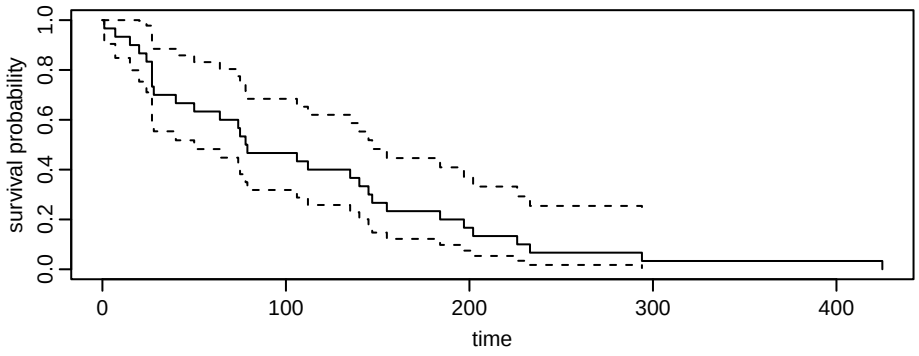
Example in R

```
t <- floor(rexp(30, rate = 1 / 100))
plot(sort(t), xlab = "id", ylab = "time survived")
```



Example in R

```
library(survival)
plot(survfit(Surv(t, rep(1, 30)) ~ 1),
     xlab = "time",
     ylab = "survival probability")
```



The Kaplan–Meier estimator is designed to handle censoring from above.

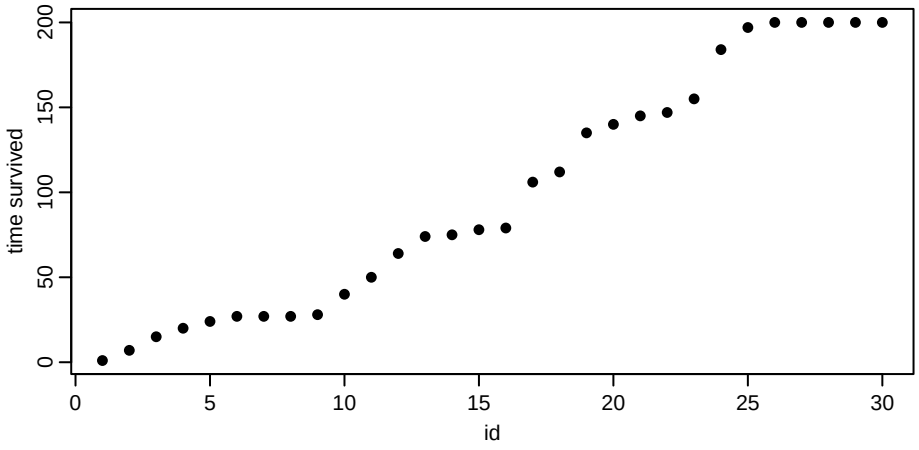
Typically, there is some maximum observable time t_*.

That is, if $T > t_*$ we cannot observe it's value, only that it's greater than t_*.

For example, a patient survived to the end of the study.

Example in R

```
 t <- pmin(t, 200) # Take the minimum of t and t_star = 200
plot(sort(t), ylab = "time survived", xlab = "id")
```

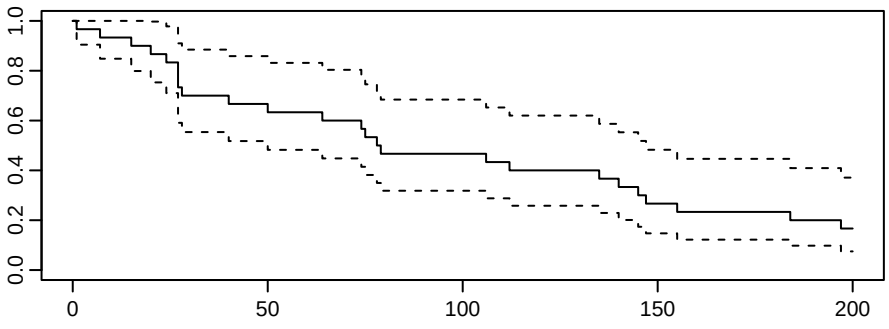


Example in R

```
observed <- (t != 200)
Surv(t, observed)
```

```
##  [1]  64   79  112   40    7  184   74   24  155  106   27  200+ 200+  50  145
## [16] 200+ 147  135   27  197   28   15   20  200+ 200+  75   78    1  140   27
```

```
plot(survfit(Surv(t, observed) ~ 1))
```



We want survival times to possibly depend on covariates.

It is common to model the *hazard function*

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

which is approximately, for small $\epsilon > 0$,

$$\frac{P(T < t + \epsilon \mid T \geq t)}{\epsilon}.$$

Example

Constant hazard characterizes the exponential distribution:

The cdf and pdf of the exponential with mean $1/\lambda$ are, respectively, $F(t) = 1 - \exp(-\lambda t)$ and $f(t) = \lambda \exp(-\lambda t)$. Therefore,

$$\frac{f(t)}{S(t)} = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda.$$

The Cox proportional hazards regression model assumes

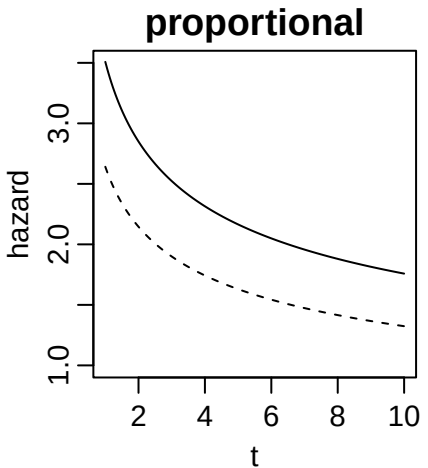
$$h(t; X_i) = h_0(t) \exp\left(\sum_{j=1}^{p} X_{ij} \beta_j \right),$$

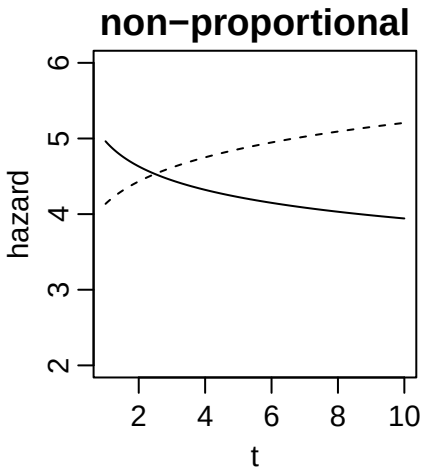
where h_0 is called the baseline hazard function.

It is the hazard function for someone with all $X_{ij} = 0$ ("intercept").

- $h_0(t) = \lambda$ for all t and $\beta_j = 0$ for all j corresponds to the exponential distribution for T.

The hazards are proportional in the sense that the ratio of the hazards for two different covariate vectors do not depend on t.





One can show that

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

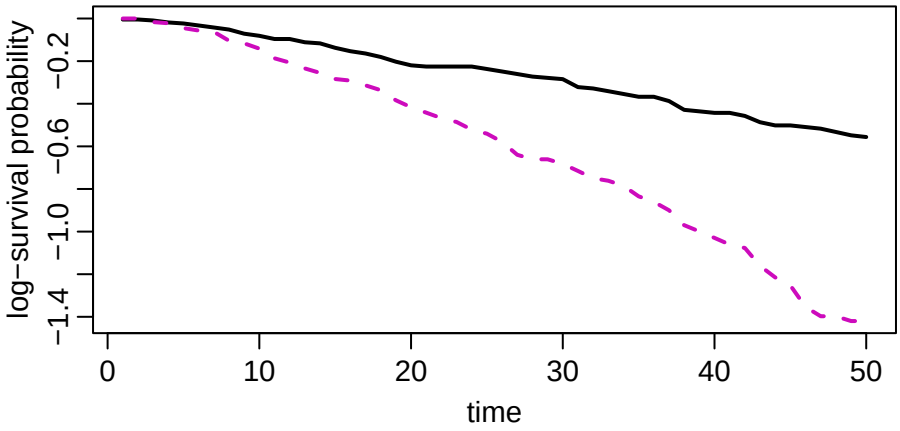
Thus, two hazards h_1 and h_2 are proportional if

$$c\log\{S_1(t)\} = \log\{S_2(t)\}.$$

for some constant c. Loosely speaking, plots of estimated log-survival functions should have similar shapes.

The quantity $-\log S(t)$ is called the cumulative hazard because it is equal to $\int_0^t h(s)ds$.

Proportional hazards example



Fitting a Cox proportional hazards model

```
fit <- survival::coxph(Surv(time_to_find, find_cheese) ~ long_training, data = lab5_dat)
summary(fit)
```

```
## Call:
## survival::coxph(formula = Surv(time_to_find, find_cheese) ~ long_training,
##     data = lab5_dat)
##
##   n= 400, number of events= 252
##
##                     coef exp(coef) se(coef)     z Pr(>|z|)
## long_trainingTRUE 0.9401    2.5603   0.1297 7.248 4.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## long_trainingTRUE      2.56     0.3906     1.986     3.301
##
## Concordance= 0.608  (se = 0.016 )
## Likelihood ratio test= 53.8  on 1 df,   p=2e-13
## Wald test            = 52.53  on 1 df,   p=4e-13
```

Plotting the estimated baseline cumulative hazard

```
bh <- survival::basehaz(survival::coxph(Surv(time_to_find, find_cheese) ~ long_training,
                                        data = lab5_dat),
            centered = F)
plot(bh$hazard ~ bh$time, type = "l", lwd = 2, xlab = "time", ylab = "cumulative hazard")
```