MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 7, spring 2019

Logistic regression, probabilities, odds and odds ratio

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

16/4-19

Introduction to Logistic regression

- In this part of the course we consider a *nonlinear model* (nonlinear in the parameters).
- This time our response variable Y will be a **discrete**, **binary variable** (success/failure, yes/no, etc).
- The nature of the response will make the Bernoulli (a special case of the Binomial) distribution a natural choice.
- The resulting regression model is called **logistic regression**.
- Our expected response will be the probability of success.

Why is this relevant?

Examples:

- political election: response is win/lose. What factors (covariates) affect the probability to win? (e.g. money spent on campaign; age of the candidate etc.)
- result of some medical test (positive/negative): estimate the probability to have a "positive" result, depending on several physiological covariates.
- crash test dummies. Probability of "survival" of a dummy, depending on several test conditions.
- ...

We consider logistic regression with binary response. But extension to *multicategory* (or polytomous) response are possible, assuming a multinomial distributed response.

Binomial distribution

Let Y be the number of successes in n independent trials, each with the same probability of success, p. Then $Y \sim Bin(n, p)$ with

$$Pr(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

$$E(Y) = np, \qquad Var(Y) = np(1-p).$$

For the estimate $\hat{p} = Y/n$ we have

$$\hat{p} \approx N(p, \frac{p(1-p)}{n}) \qquad I_p = (\hat{p} \pm \lambda_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

when n is "large enough", typically when $np(1-p) > 10$.
Warning: If n is too small the interval can go outside $[0, 1]$.

Covariates

Before:

Y_i was a continuous variable with

$$Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \Leftrightarrow$$
$$Y_i \sim N(\mu_i, \sigma^2)$$
$$E(Y_i) = \mu_i = \mathbf{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

Now:

Y_i is discrete with two possible outcomes: success (=1) or failure (=0) with probabilities $Pr(Y_i = 1) = p_i$ and $Pr(Y_i = 0) = 1 - p_i$

$$Y_i \sim Bin(1, p_i) \text{ with } Pr(Y_i = k) = p_i^k(1 - p_i)^{1-k},\ k = 0, 1$$
$$E(Y_i) = p_i = \text{ some function of } \mathbf{x}_i$$
$$V(Y_i) = p_i(1 - p_i) \text{ depends on } \mathbf{x}_i$$

Choosing $p_i = \mathbf{x}_i\boldsymbol{\beta}$ is *not* good since $0 \leq p_i \leq 1$.

Odds

The odds of "success" is defined as

$$\begin{aligned}\text{Odds} &= \frac{P(\text{success})}{P(\text{failure})} = \frac{p}{1-p} \\ &= \text{number of successes for each failure}\end{aligned}$$

$$\text{log-odds} = \ln \text{Odds} = \ln \frac{p}{1-p} = \text{logit}(p)$$

$$\text{Odds}_{\text{failure}} = \frac{1}{\text{Odds}_{\text{success}}},$$

$$\ln \text{Odds}_{\text{failure}} = -\ln \text{Odds}_{\text{success}}$$

	min	middle	max	
p	0	1/2	1	
Odds	0	1	∞	
ln Odds	$-\infty$	0	∞	no bounds!

Logistic regression model

$$
\begin{aligned}
Y_i &= \text{"success"}\ (=1) \text{ or "failure"}\ (=0) \\
P(Y_i = 1) &= 1 - P(Y_i = 0) = p_i \\
Y_i &\sim Bin(1, p_i), \quad i = 1, \ldots, n, \ \text{(independent)} \\
\text{logit}(p_i) &= \ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} = \mathbf{x}_i \boldsymbol{\beta} \\
p_i &= \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i \boldsymbol{\beta}}}
\end{aligned}
$$

Parameters

$\beta_0 =$ log-odds and $e^{\beta_0} =$ odds when all X_{ji} are 0,

$\beta_j =$ additive change in log-odds and...

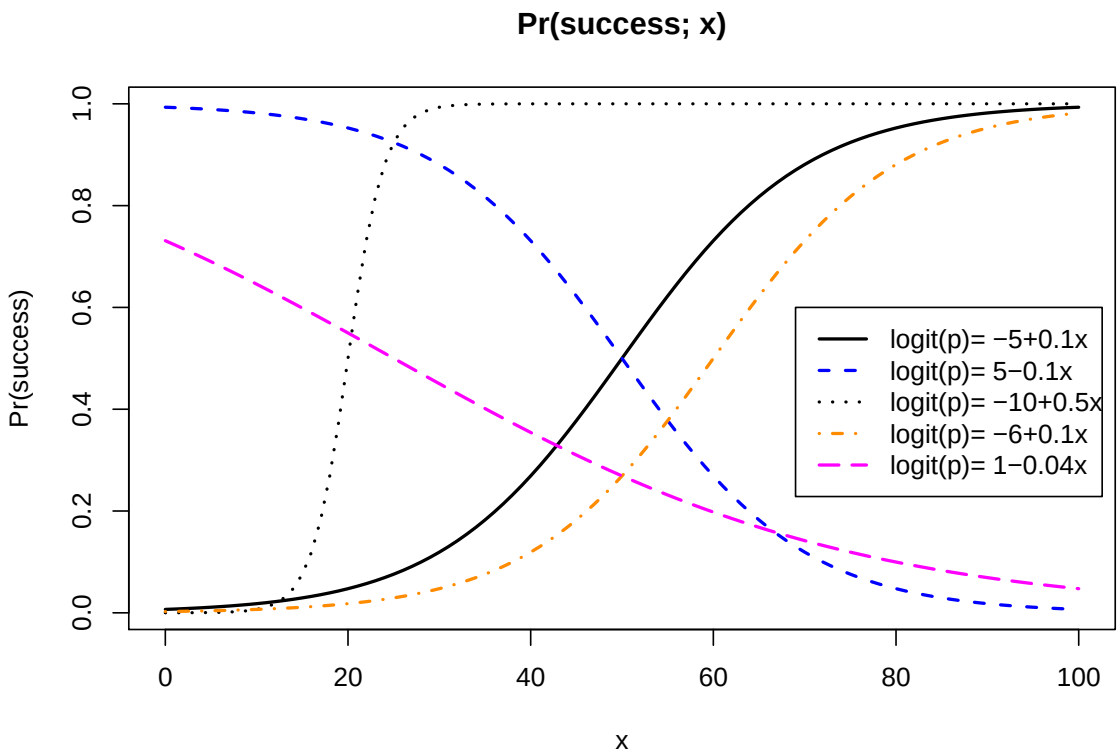
$e^{\beta_j} =$ relative change in odds when X_{ji} is increased by 1, $j = 1, \ldots, p$

Logistic regression with one continuous covariate

Simple logistic regression model:

$$Y_i = \text{"success"}\ (=1) \text{ or } \text{"failure"}\ (=0) \in Bin(1, p_i)$$

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_i, \qquad p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$



Interpretation of β_1

- What happens when we increase X by 1?

$$OR = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

If $\beta_1 = 0.04$ then $e^{\beta_1} = 1.04$ and the odds increases by 4 %.
If $\beta_1 = -0.04$ then $e^{\beta_1} = 0.96$ and the odds decreases by 4 %.

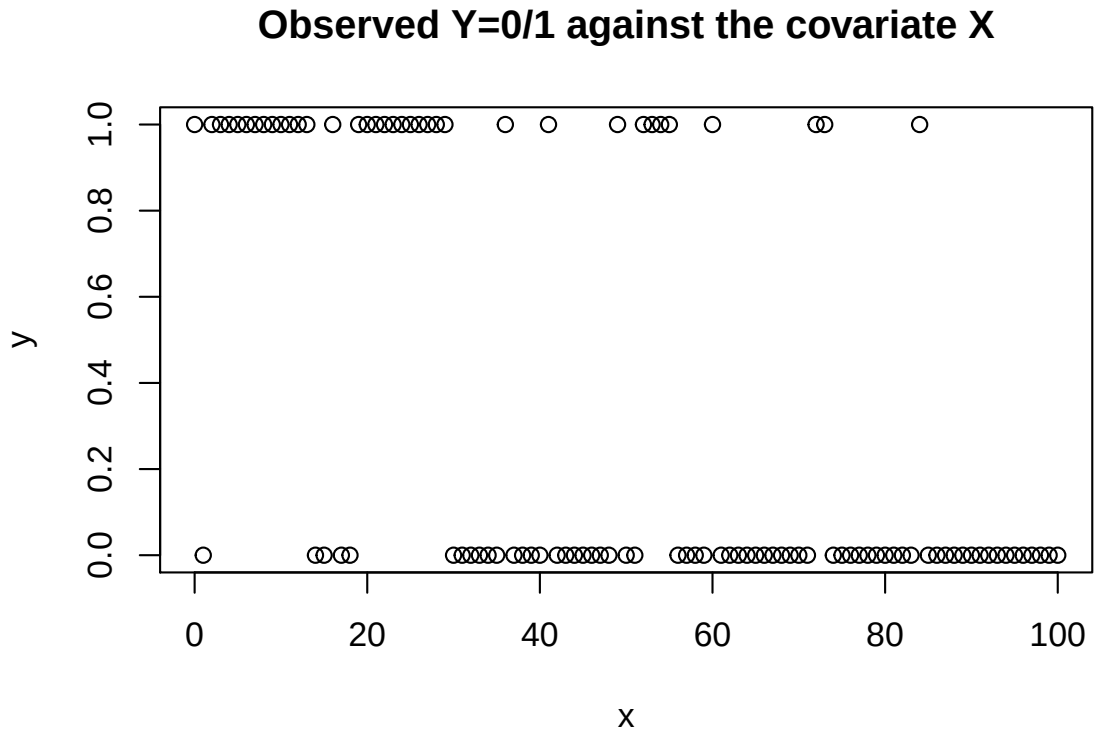
- What happens when we increase X by 10?

$$OR = \frac{e^{\beta_0+\beta_1(x+10)}}{e^{\beta_0+\beta_1 x}} = e^{10\beta_1} = (e^{\beta_1})^{10}$$

If $\beta_1 = 0.04$ then $(e^{\beta_1})^{10} = 1.04^{10} = 1.49$ and the odds increases by 49 %.
If $\beta_1 = -0.04$ then $(e^{\beta_1})^{10} = 0.96^{10} = 0.67$ and the odds decreases by 33 %.

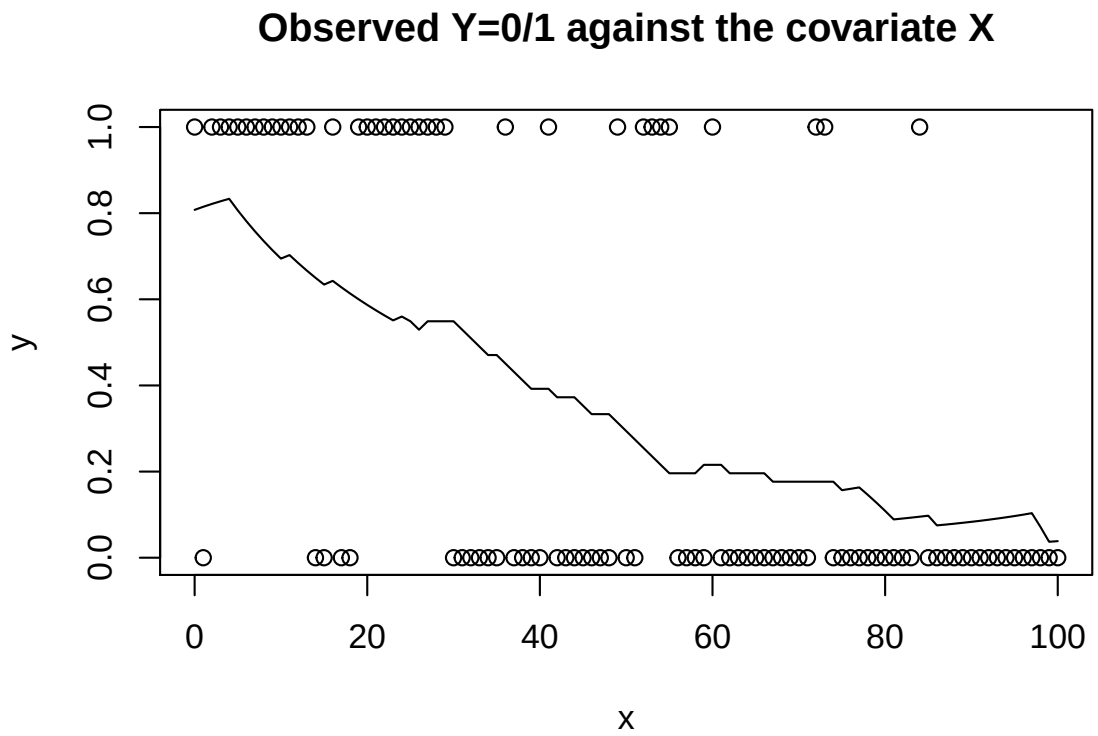
Does the data follow an S-shape?

In simple linear regression we plotted Y against X to see if it looked like a straight line. What about now?



Well...

We can get a rough estimate of the shape using a weighted moving average, e.g. a Kernel smoothing regression which calculates the average Y-value in an interval moving along the x-axis. The width of the interval is decided by the bandwidth. Experiment until you get something reasonably smooth.



Sort of S-shaped. Obviously $\beta_1 < 0$.

Using R

Use `glm` (Generalized Linear Model) with `family="binomial"`:

```
model <- glm(green ~ type, family = "binomial",
             data = fruit)
```

Prediction:

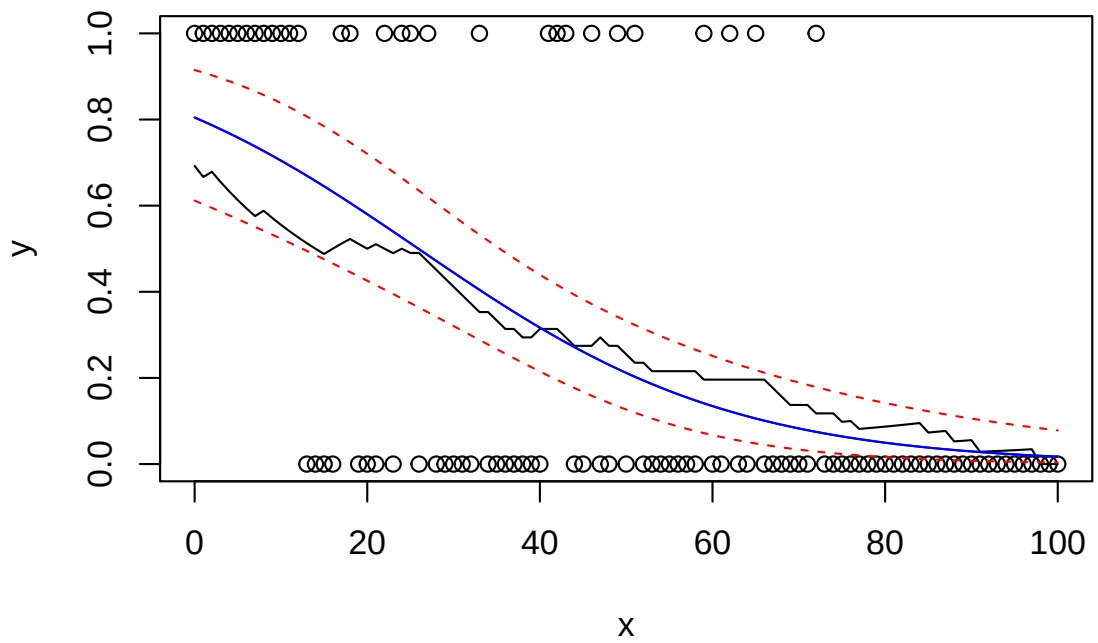
- Using `predict` by default returns log(odds)

  ```
  logodds <- predict(model, x0)
  ```

  $\ln(\frac{\hat{p}_0}{1-\hat{p}_0}) = \mathbf{x}_0\hat{\boldsymbol{\beta}}$
- Do you want the estimated probabilities instead?

  ```
  prob <- predict(model, x0, type = "response")
  ```

  $\hat{p}_0 = \frac{e^{\mathbf{x}_0\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_0\hat{\boldsymbol{\beta}}}}$




```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.81148    0.50712   3.572 0.000354 ***
x           -0.05447    0.01111  -4.903 9.45e-07 ***

beta           2.5%  97.5%    exp(beta)     2.5%  97.5%
(Intercept)  0.496  2.437    (Intercept)  1.642  11.44
x           -0.080 -0.034    x            0.923  0.967
```

Confidence intervals for (log) odds ratios

We will show later (Lecture 8) that the $\hat{\beta}_j$ are asymptotically normally distributed. Thus, we can construct confidence intervals in the usual way (define λ_α as the α-percentile from N(0,1)):

$$I_{\ln OR_j} = I_{\beta_j} = (\hat{\beta}_j \pm \lambda_{\alpha/2} \cdot SE(\hat{\beta}_j))$$

then exponentiate the bounds

$$I_{OR_j} = I_{e^{\beta_j}} = e^{I_{\beta_j}} = (e^{\hat{\beta}_j - \lambda_{\alpha/2} \cdot SE(\hat{\beta}_j)}, \; e^{\hat{\beta}_j + \lambda_{\alpha/2} \cdot SE(\hat{\beta}_j)})$$

Test using intervals

- If I_{β_j} contains 0 then variable X_j is not significant.
- If $I_{e^{\beta_j}}$ contains $e^0 = 1$ then variable X_j is not significant.

Wald test for β_j

Does variable X_j have a significant effect on the probability of success, i.e., does it change the odds of success?

Wald test

We want to test H_0: $\beta_j = 0$ against H_1: $\beta_j \neq 0$. If H_0 is true then

$$Z = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \approx N(0,1) \qquad \text{if } n \text{ is large}$$

and we should reject H_0 at significance level α if

$$\frac{|\hat{\beta}_j - 0|}{SE(\hat{\beta}_j)} > \lambda_{\alpha/2}$$

Later we will show that the $\hat{\beta}_j$ are maximum likelihood estimates which are (asymptotically as $n \to \infty$) Normally distributed.

Why a "Wald test"?

A "Wald test" test is constructed similarly to a t-test but....it's not distributed as a Student's t (hence a different name).

Asymptotic normality of a MLE

Under fairly weak regularity conditions the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is asymptotically Normal. Hence $Z \sim N(0,1)$ as $n \to \infty$ and we can use the $\lambda_{\alpha/2}$-quantiles.

This is why in the `summary()` we get `z value` (not `t value`):

Probability estimates

Since the log odds is a linear function

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = \mathbf{x}_i \boldsymbol{\beta}$$

the corresponding probability of success becomes

$$p_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_p X_{pi}}}{1+e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_p X_{pi}}} = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}$$

which is a non-linear function of the β-parameters.
Since $\mathbf{x}_i\hat{\boldsymbol{\beta}}$ is a linear function of (dependent) normally distributed β-estimates we can construct confidence intervals for the log odds:

$$I_{\mathbf{x}_i\boldsymbol{\beta}} = (\mathbf{x}_i\hat{\boldsymbol{\beta}} \pm \lambda_{\alpha/2} \cdot SE(\mathbf{x}_i\hat{\boldsymbol{\beta}}))$$

Since $\hat{p}_i$ is a monotonous, increasing, function of $\mathbf{x}_i\hat{\boldsymbol{\beta}}$ we get

$$I_{p_i} = \frac{e^{I_{\mathbf{x}_i\boldsymbol{\beta}}}}{1+e^{I_{\mathbf{x}_i\boldsymbol{\beta}}}} \qquad \text{which always lies is } [0,1]!$$

Example with a single categorical covariate

We have a basket full of green and yellow fruit. We want to model the probability to **pick a green fruit**.

- $p_i = Pr(i\text{:th picked fruit is green})$ depending on fruit type.
- which fruit type is more likely to be green? Or, are some fruits more likely to be green than others, or is color unrelated to type of fruit?
- Odds ratios will help...
- notice: response is categorical. Covariate is categorical. Hence data can be cross-tabled:

Fruit	Green	Yellow	Total	$\hat{p}_i$	$\widehat{odds}_i$	$\widehat{OR}_i$
Banana (ref.)	100	200	300	1/3	1/2	$\frac{1/2}{1/2} \equiv 1$
Apple	40	30	70	4/7	4/3	$\frac{4/3}{1/2} = 8/3$
Melon	20	20	40	1/2	1	$\frac{1}{1/2} = 2$
Total	160	250	410			

Of course a categorical covariate is modeled using dummy variables. So the previous example becomes

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \qquad \text{for } i = 1, \ldots, n$$

where $X_{1i} = 1$ for Apples and 0 otherwise, $X_{2i} = 1$ for Melons and 0 otherwise. Then

$$\text{log-odds}_i = \ln \frac{p_i}{1-p_i} = \begin{cases} \beta_0 & \text{Banana (reference)} \\ \beta_0 + \beta_1, & \text{Apple} \\ \beta_0 + \beta_2, & \text{Melon} \end{cases}$$

$$\text{odds}_i = \frac{p_i}{1-p_i} = \begin{cases} e^{\beta_0} & \text{Banana (reference)} \\ e^{\beta_0+\beta_1} = e^{\beta_0} \cdot e^{\beta_1}, & \text{Apple} \\ e^{\beta_0+\beta_2} = e^{\beta_0} \cdot e^{\beta_2}, & \text{Melon} \end{cases}$$

Odds and odds ratios

Since change in the odds is a relative, not additive, change, it is more reasonable to compare two odds by taking their ratio, not their difference.
The odds ratios of getting a green fruit if we pick an apple or a melon, compared to if we pick a banana, are

$$\text{OR}_{\text{apple}} = \frac{\text{odds}_{\text{apple}}}{\text{odds}_{\text{banana}}} = \frac{e^{\beta_0} \cdot e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}, \qquad e^{\hat{\beta}_1} \approx 2.67$$

$$\text{OR}_{\text{melon}} = \frac{\text{odds}_{\text{melon}}}{\text{odds}_{\text{banana}}} = \frac{e^{\beta_0} \cdot e^{\beta_2}}{e^{\beta_0}} = e^{\beta_2}, \qquad e^{\hat{\beta}_2} = 2.00$$

The odds of getting a green fruit is *more than* twice as large for apples as for bananas.
The odds of getting a green fruit is twice as large for melons as for bananas.

Distribution of Odds ratios using one categorical X-variable

For small to moderate sample sizes, the distribution of the odds ratio is highly skewed! $\Rightarrow$ take (natural) logarithms and use the fact that maximum likelihood estimates are asymptotically Normal (see next lecture).

If we have a large number of observations in each of the cells then

$$\ln \hat{OR} \approx N(\ln OR, V(\ln \hat{OR}))$$

$$SE(\hat{\beta}_0) \approx \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{01}}}$$

$$SE(\ln \hat{OR}_j) = SE(\hat{\beta}_j) \approx \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{j0}} + \frac{1}{n_{j1}}}, \qquad j = 1, 2, \ldots$$

where n_{00} = number of failures in the reference group, n_{01} = number of successes in the reference group, n_{j0} = number of failures in group j etc.

Warning: Only works when all $n_{j0}, n_{j1} > 0$ so X cannot be continuous!

Intervals for the log-odds β_0 and log-odds ratios β_1 and β_2

Fruit	β_j	$\hat{\beta}_j$	$S.E.(\hat{\beta}_j)$	I_{β_j}
Banana	β_0	-0.69	0.12	$-069 \pm 1.96 \cdot 0.12 = (-0.94, -0.46)$
Apple	β_1	0.98	0.27	$0.98 \pm 1.96 \cdot 0.27 = (0.45, 1.52)$
Melon	β_2	0.69	0.34	$0.69 \pm 1.96 \cdot 0.34 = (0.02, 1.36)$

Intervals for the odds ratios e^{β_1} and e^{β_2}

Fruit	e^{β_j}	$\hat{\text{OR}}$	$C.I.(\text{OR})$
Banana (ref.)		1	$-$
Apple	e^{β_1}	$e^{0.98} = 2.67$	$(e^{0.45}, e^{1.52}) = (1.57, 4.56)$
Melon	e^{β_2}	$e^{0.69} = 2.00$	$(e^{0.02}, e^{1.36}) = (1.03, 3.91)$

The estimates of the log odds become

$$\mathbf{x}_i\hat{\boldsymbol{\beta}} = \begin{cases} \hat{\beta}_0 & = -0.69, & \text{Banana (ref.)} \\ \hat{\beta}_0 + \hat{\beta}_1 & = -0.69 + 0.98 = 0.29, & \text{Apple} \\ \hat{\beta}_0 + \hat{\beta}_2 & = -0.69 + 0.69 = 0.00, & \text{Melon} \end{cases}$$

with 95 % confidence intervals as

$$l_{\mathbf{x}_i\boldsymbol{\beta}} = \begin{cases} (-0.69 \pm 1.96 \cdot 0.12) & = (-0.93, -0.45), & \text{Banana (ref.)} \\ (0.29 \pm 1.96 \cdot 0.24) & = (-0.19, 0.76), & \text{Apple} \\ (0.00 \pm 1.96 \cdot 0.32) & = (-0.62, 0.62), & \text{Melon} \end{cases}$$

The standard errors are calculated using the covariance matrix for $\hat{\boldsymbol{\beta}}$, see Lecture 8.

The estimates of the probabilities become

$$\hat{p}_i = \begin{cases} \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} & = \frac{e^{-0.69}}{1+e^{-0.69}} \approx 0.33, \quad \text{Banana (ref.)} \\ \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} & = \frac{e^{0.29}}{1+e^{0.29}} \approx 0.57, \quad \text{Apple} \\ \frac{e^{\hat{\beta}_0+\hat{\beta}_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_2}} & = \frac{e^{0.00}}{1+e^{0.00}} = 0.50, \quad \text{Melon} \end{cases}$$

with 95 % confidence intervals as

$$I_{p_i} = \begin{cases} \left(\frac{e^{-0.93}}{1+e^{-0.93}}, \frac{e^{-0.45}}{1+e^{-0.45}}\right) & = (0.28, 0.39), \quad \text{Banana (ref.)} \\ \left(\frac{e^{-0.19}}{1+e^{-0.19}}, \frac{e^{0.76}}{1+e^{0.76}}\right) & = (0.45, 0.68), \quad \text{Apple} \\ \left(\frac{e^{-0.62}}{1+e^{-0.62}}, \frac{e^{0.62}}{1+e^{0.62}}\right) & = (0.35, 0.65), \quad \text{Melon} \end{cases}$$

Appendix: Proof for the SE of the Odds (*)

***[For the interested reader only: not an exam topic]**

This requires first introducing the "delta method" for obtaining standard errors. Assume $\hat{\theta}$ is the estimator of a parameter θ (you might think that $\theta \equiv \beta_j$ and $\hat{\theta} \equiv \hat{\beta}_j$). Also assume that $\hat{\theta}$ is asymptotically Gaussian.

- from the assumptions:

$$\frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \to N(0,1) \qquad (n \to \infty)$$

- the "delta method" says that for a non-zero valued function $f(\theta)$ having first derivative $f'(\theta)$ (> 0, else change sign), we have

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})SE(\hat{\theta})} \to N(0,1) \qquad (n \to \infty)$$

- therefore the asymptotic mean of $f(\hat{\theta})$ is $f(\theta)$ and the asymptotic SE of $f(\hat{\theta})$ is estimated with $f'(\hat{\theta})SE(\hat{\theta})$.

Motivation for the delta method

- if $\hat{\theta}$ is close to θ

$$\frac{f(\hat{\theta}) - f(\theta)}{\hat{\theta} - \theta} \approx f'(\hat{\theta})$$

- so

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})} \approx \hat{\theta} - \theta$$

- therefore

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})SE(\hat{\theta})} \approx \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

We want to compute $SE(\hat{\beta}_0) \equiv SE(\ln odds)$ (you can follow a similar procedure for $SE(\ln OR_j)$). To simplify things let's assume a 2x2 table:

	success	fail	tot	p	odds
category 0	n_{01}	n_{00}	n_0	n_{01}/n_0	n_{01}/n_{00}
category 1	n_{11}	n_{10}	n_1	n_{11}/n_1	n_{11}/n_{10}

Now let's focus on category 1: over n_0 independent experiments we got n_{01} successes, each with probability $\hat{p}_1 = n_{01}/n_0$. Hence we write that n_{01} is the outcome of a binomial distribution, $n_{01} \sim Bin(n_0, p_1)$ while $n_{00} = n_0 - n_{01}$. Note that n_0 is fixed. The variance of a generic $Bin(n, p)$ distribution is $np(1-p)$ hence

$$\hat{V}(n_{01}) = n_0 \hat{p}_1 (1 - \hat{p}_1) = n_0 \cdot \frac{n_{01}}{n_0} \cdot \frac{n_0 - n_{01}}{n_0} = \frac{n_{01}(n_0 - n_{01})}{n_0}$$

Let's take $\ln \text{odds}_1 = \ln n_{01} - \ln n_{00} = \ln n_{01} - \ln(n_0 - n_{01})$ and we have

$$V(\ln \text{odds}_1) = V(\ln n_{01} - \ln(n_0 - n_{01})) = V(f(n_{01}))$$

where we use the delta method with

$$f(x) = \ln x - \ln(n_0 - x),$$
$$f'(x) = \frac{1}{x} + \frac{1}{n_0 - x} = \frac{n_0}{x(n_0 - x)}$$

This gives us

$$SE(\ln \text{odds}_1) = SE(f(n_{01})) \approx f'(n_{01}) SE(n_{01})$$
$$= \frac{n_0}{n_{01}(n_0 - n_{01})} \sqrt{\frac{n_{01}(n_0 - n_{01})}{n_0}} = \sqrt{\frac{n_0^2 n_{01}(n_0 - n_{01})}{n_0 n_{01}^2 (n_0 - n_{01})^2}}$$
$$= \sqrt{\frac{n_0}{n_{01}(n_0 - n_{01})}} = \sqrt{\frac{1}{n_{01}} + \frac{1}{n_0 - n_{01}}} = \sqrt{\frac{1}{n_{01}} + \frac{1}{n_{00}}}$$

Similar for $\ln \text{odds}_2$ and then using that the two odds are independent gives the SE for the log odds ratio as well.