Logistic regression

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Introduction to Statistics

From linear to logistic regression

- In the previous lecture we used linear regression to estimate the effect of alcohol intake on breast density
- Linear regression is suitable for continuous outcomes (e.g. breast dense volume)
- For binary outcomes (e.g. breast cancer), it is more common to use logistic regression

Outline

Outline

Alcohol intake vs breast cancer

- We saw previously that alcohol seems to *decrease* the risk of breast cancer
  - possibly explained by confounding by age; we address this later
- To motivate simple logistic regression we will explore this relation further
- Specifically we will aim to answer the following research question:

How much does the breast cancer risk decrease, on average, with each additional gram alcohol intake?

- Can we not use linear regression to answer this question?

Fitting a linear regression model

- Suppose we fit the linear regression model

$$\mu_{y|x} = \alpha + \beta x$$

  with x = alcohol intake and y = breast cancer (0/1)
- We then get

```
> m <- glm(formula=bc~alc, data=bc)
> summary(m)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3250047  0.0124934   26.01   <2e-16 ***
alc         -0.0151287  0.0009249  -16.36   <2e-16 ***
```

- *How can we interpret the estimated parameters?*

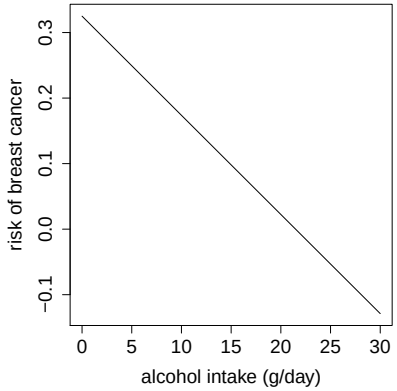
Solution

$$\mu_{y|x} = \alpha + \beta x$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3250047  0.0124934   26.01   <2e-16 ***
alc         -0.0151287  0.0009249  -16.36   <2e-16 ***
```

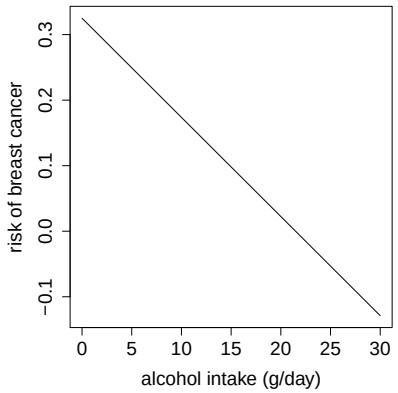
- For binary (0/1) variables, the mean is the proportion of 1's
  - e.g. the proportion of breast cancer diagnoses - the risk of breast cancer - during follow-up
- α is the risk of breast cancer for those who don't drink any alcohol
  - estimated risk = 33%
- β is the difference in risk of breast cancer between two groups who differ 1 g/day in alcohol intake
  - estimated risk reduction = 1.5 percentage points per g/day

Risk of breast cancer, according to the model



- *Do you see any problem with this model?*

Solution



- According to the model, the risk of breast cancer is negative for alcohol intakes $>$ 21.8 g/day
- This is not logically possible

Outline

The odds

- A risk (chance/proportion) is restricted to the range $(0, 1)$
- Restrictions are inconvenient in regression models; it is easier to work with unrestricted parameters
- For this purpose we shift focus and consider the odds instead

The odds (recap)

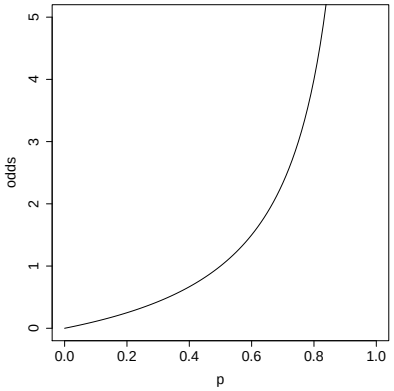
- Let p be the risk (proportion) that we are interested in
  - e.g. the risk of breast cancer during follow-up
- The odds is defined as

$$\text{odds} = \frac{p}{1-p}$$

- It measures how much more common it is to have the outcome, than to not have the outcome
  - e.g. if 2/3 of all subjects are sick, then the odds of being sick is (2/3)/(1/3)=2; twice as likely to be sick as non-sick
- Can calculate the risk from the odds as

$$p = \frac{\text{odds}}{1+\text{odds}}$$

Risk vs odds (recap)



- When the risk is small, the odds is approximately equal to the risk
- The odds is restricted to the range $(0, \infty)$

The log odds

- The odds is less restricted than the risk, but not completely unrestricted
- We consider the logarithm of the odds - the log odds

$$\log \text{odds} = \log\left(\frac{p}{1-p}\right)$$

- Also written as $\text{logit}(p)$

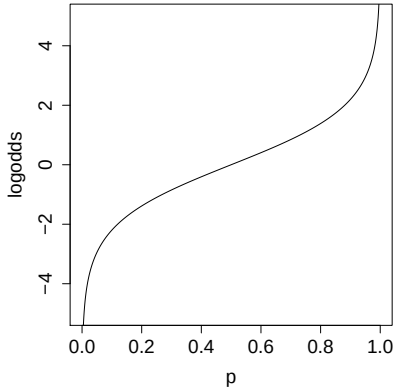
A note on terminology

- In statistics,'log' is the natural logarithm

$$e^{\log \text{odds}} = \text{odds}$$

  where e is Euler's number = 2.718...
- For instance, $\log \text{odds} = 1$ means that $\text{odds} = e^1 = 2.718..$
- In other branches of science, the natural logarithm is often denoted with 'ln'

Relation between the risk and the log odds



- The log odds is unrestricted

Relation between the risk and the log odds, cont'd

- The risk and the log odds contain the same information, and one can always be compute from the other

$$\begin{aligned} \text{log odds} &= \log\left(\frac{p}{1-p}\right) \\ p &= \frac{e^{\text{log odds}}}{1+e^{\text{log odds}}} \end{aligned}$$

- However, since the log odds is unrestricted it is more convenient to use in regression models than the risk

The simple logistic regression model

- One outcome y (breast cancer) and one covariate x (alcohol intake)
- $p_{y|x}$ is the risk of y for those subjects with a specific value of x
  - e.g. $p_{y|x=10}$ is the risk of breast cancer for women with an alcohol intake of 10 g/day
- The log odds of y is assumed to be a linear function of x

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

- Equivalent formulation for the risk

$$p_{y|x} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Terminology

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

- y is called outcome, regressand, endogenous variable, response variable, or dependent variable
- x is called exposure, covariate, predictor, regressor, exogenous variable, explaining variable, or independent variable
  - with several covariates in the model (more later), the term 'exposure' usually refers to the covariate of special interest
- α is called intercept
- β is called slope

Interpretation of α

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

$$\text{logit}(p_{y|x=0}) = \alpha + \beta \times 0 = \alpha$$

- α is the log odds of y for those with $x = 0$
  - e.g. the log odds of breast cancer for those who don't drink any alcohol

Interpretation of β

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

$$\text{logit}(p_{y|x+1}) - \text{logit}(p_{y|x}) = \{\alpha + \beta(x+1)\} - \{\alpha + \beta x\} = \beta$$

- β is the difference in the log odds of y between two groups who differ with 1 unit in x
  - e.g. the difference in log odds of breast cancer between two groups who differ 1 g/day in alcohol intake
- The association between x and y
- If $\beta = 0$, then the log odds of y does not depend on x
  - and neither does the risk

Fitting the logistic regression model in R

- A method called 'maximum likelihood'
  - generalization of least squares, minimizes the 'distance' between the fitted model and the observed data

```
> m <- glm(formula=bc~alc, family="binomial",
  data=bc)
> summary(m)

             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12439    0.10902  -1.141    0.254
alc         -0.15451    0.01003 -15.406   <2e-16 ***
```

- *Interpretation?*
- *Plot the log odds of breast cancer as a function of alcohol intake*
- *Estimate the risk of breast cancer for women who drink 0, 10 and 20 g/day*
- *Plot the risk of breast cancer as a function of alcohol intake*

Solution

$$\mu_{y|x} = \alpha + \beta x$$

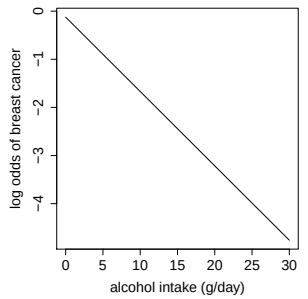
```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12439    0.10902  -1.141    0.254
alc         -0.15451    0.01003 -15.406   <2e-16 ***
```

- Those who did not drink any alcohol at enrollment had a log odds of breast cancer during follow-up equal to -0.12
- For two groups who differed in alcohol intake with 1 g/day at enrollment, the difference in log odds of breast cancer was equal to -0.15

Solution, cont'd

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12439    0.10902  -1.141    0.254
alc         -0.15451    0.01003 -15.406   <2e-16 ***
```



Solution, cont'd

$$p_{y|x} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12439    0.10902  -1.141    0.254
alc         -0.15451    0.01003 -15.406   <2e-16 ***
```

$$\hat{p}_{y|x=0} = \frac{e^{\hat{\alpha}+\hat{\beta}\times 0}}{1+e^{\hat{\alpha}+\hat{\beta}\times 0}} = 0.47$$

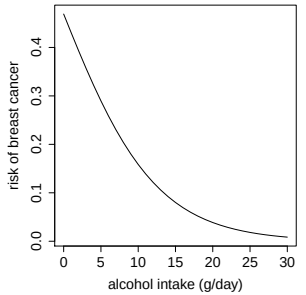
$$\hat{p}_{y|x=10} = \frac{e^{\hat{\alpha}+\hat{\beta}\times 10}}{1+e^{\hat{\alpha}+\hat{\beta}\times 10}} = 0.16$$

$$\hat{p}_{y|x=20} = \frac{e^{\hat{\alpha}+\hat{\beta}\times 20}}{1+e^{\hat{\alpha}+\hat{\beta}\times 20}} = 0.04$$

Solution, cont'd

$$p_{y|x} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12439    0.10902  -1.141    0.254
alc         -0.15451    0.01003 -15.406   <2e-16 ***
```



Confidence intervals for the model parameters

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

```
> confint(m)

                2.5 %     97.5 %
(Intercept) -0.3381222  0.0893226
alc         -0.1744023 -0.1350786
```

When are the CI's and p-values valid?

- Due to the CLT, the CI's and p-values are roughly correct if the sample is large
- If the sample is small, then the CI's and p-values may be very wrong

Outline

Linearity assumption

$$\text{logit}(p_{y|x}) = \alpha + \beta x$$

- The logistic regression model assumes that the log odds of y is a linear function of x
- We must verify that the model fits reasonably well to the data
- This is usually done with a residual plot

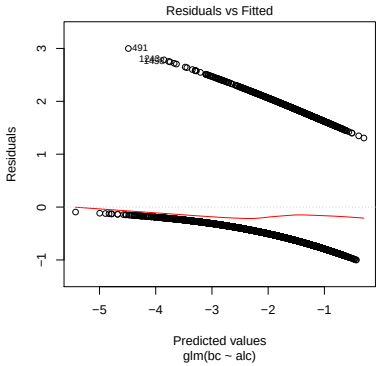
Residuals

- The residuals for logistic regression are usually defined somewhat differently than for linear regression
  - exact definition beyond the scope of this course
- Same idea though
  - the residuals measure the 'distance' between the fitted model and the observed data
  - if the model fits well, then the residuals should have roughly mean 0 everywhere

Residual plot in R

```
> m <- glm(formula=dens~alc, family="binomial",
  data=bc)
> plot(m, which=1)
```

- Residual vs $\text{logit}(p_{y|x}) = \hat{\alpha} + \hat{\beta} x$



- Looks quite ok

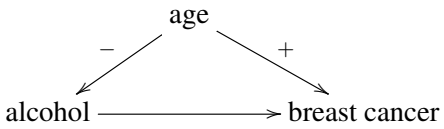
Outline

Association = causation?

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12439    0.10902  -1.141    0.254
alc         -0.15451    0.01003 -15.406   <2e-16 ***
```

- We have used logistic regression to establish an inverse statistical association between alcohol intake and breast cancer
- *Does this mean that alcohol actually decreases the risk of breast cancer?*

Association $\neq$ causation!



- A possible non-causal explanation:
  - young women drink more alcohol than old women
  - young women have lower risk of breast cancer than old women
  - thus; those who have a high alcohol intake tend to be young, and tend therefore to have a low risk of breast cancer
- Confounding by age; we will attempt to solve this problem with multiple logistic regression

The multiple logistic regression model

- One outcome y (breast cancer) and two covariates x (alcohol intake) and z (age)
  - can have more than two covariates as well
- $p_{y|x,z}$ is the risk of y for those subjects with a specific level of x and z
  - e.g. $p_{y|x=10,z=20}$ is the risk of breast cancer for women who have an alcohol intake of 10 g/day and are 20 years old
- The log odds of y is assumed to be a linear function of x and z:

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

Terminology

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

- Both x and z are called covariates
  - or regressors, exogenous variables, explaining variables, or independent variables
- The term 'exposure' is reserved for the covariate of main interest; x (alcohol intake) in our case

Interpretation of α

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

$$\text{logit}(p_{y|x=0,z=0}) = \alpha + \beta \times 0 + \gamma \times 0 = \alpha$$

- α is the log odds of y for those with $x = 0$ and $z = 0$
  - e.g. the log odds of breast cancer for those who don't drink any alcohol and are newborn (not very meaningful)

Interpretation of β

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

$$\text{logit}(p_{y|x+1,z}) - \text{logit}(p_{y|x,z}) = \{\alpha + \beta(x+1) + \gamma z\} - \{\alpha + \beta x + \gamma z\} = \beta$$

- β is the difference in the log odds of y between two groups who differ by 1 unit in x, but have the same level of z
  - e.g. the difference in the log odds of breast cancer between two groups who differ by 1 g/day in alcohol intake, but have the same age
- The association between x and y, adjusted for z

Interpretation of γ

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

$$\text{logit}(p_{y|x,z+1}) - \text{logit}(p_{y|x,z}) = \{\alpha + \beta x + \gamma(z+1)\} - \{\alpha + \beta x + \gamma z\} = \gamma$$

- γ is the difference in the log odds of y between two groups who differ by 1 unit in z, but have the same level of x
  - e.g. the difference in the log odds of breast cancer between two groups who differ by 1 year in age, but have the same alcohol intake
- The association between z and y, adjusted for x

Fitting the multiple logistic regression model in R

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

```
> m <- glm(formula=bc~alc+age, family="binomial",
  data=bc)
> summary(m)

             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.383773   0.555325  -9.695   <2e-16 ***
alc          0.020680   0.019945   1.037      0.3
age          0.066855   0.006927   9.652   <2e-16 ***
```

- *Interpretation?*
- *Draw a plot that illustrates how the log odds of breast cancer depends on alcohol intake, for those who are 20 years and for those who are 50 years, respectively*

Solution

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.383773   0.555325  -9.695   <2e-16 ***
alc          0.020680   0.019945   1.037      0.3
age          0.066855   0.006927   9.652   <2e-16 ***
```

- The log odds of breast cancer for those who don't drink any alcohol and are newborn, is -5.38 (not very meaningful)
- The difference in log odds of breast cancer between two groups who differ by 1 g/day in alcohol intake, but have the same age, is 0.021
  - not statistically significant
- The difference in log odds of breast cancer between two groups who differ by 1 year in age, but have the same alcohol intake, is 0.067
  - statistically significant

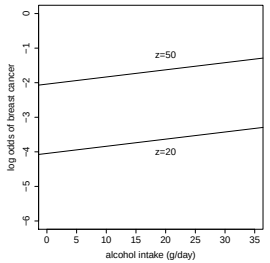
Solution, cont'd

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$

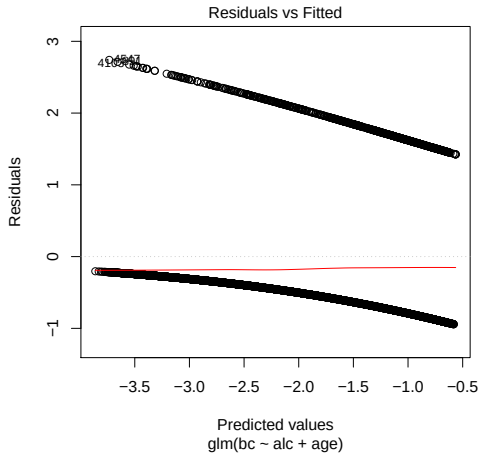
```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.383773   0.555325  -9.695   <2e-16 ***
alc          0.020680   0.019945   1.037      0.3
age          0.066855   0.006927   9.652   <2e-16 ***
```

$$\text{logit}\left(\hat{\mu}_{y|x,z=20}\right) = \hat{\alpha} + \hat{\beta}x + \hat{\gamma} \times 20 = -4.05 + 0.021x$$

$$\text{logit}\left(\hat{\mu}_{y|x,z=50}\right) = \hat{\alpha} + \hat{\beta}x + \hat{\gamma} \times 50 = -2.04 + 0.021x$$



Residual plot

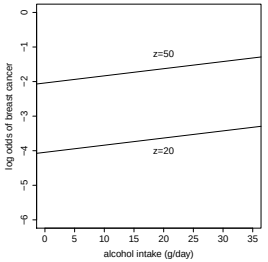


- Looks quite ok

Outline

No-interaction assumption

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z$$



- The model we have used assumes that the association between alcohol intake and breast cancer does not depend on age
  - no interactions
- To relax the no-interaction assumption, we can fit a model that includes an interaction term between x and z

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z + \psi xz$$

Fitting the model with interaction in R

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z + \psi xz$$

```
> m <- glm(formula=dens~alc+age+alc*age, data=bc)
> summary(m)

(Intercept) -4.4406472  0.6923502  -6.414 1.42e-10 ***
alc         -0.0622077  0.0416794  -1.493   0.1356
age          0.0482226  0.0106324   4.535 5.75e-06 ***
alc:age      0.0018085  0.0007812   2.315   0.0206 *
```

- The interaction term is statistically significant, but is it large enough to be relevant?
- *Draw a plot that illustrates how the log odds of breast cancer depends on alcohol intake, for those who are 20 years and for those who are 50 years, respectively*

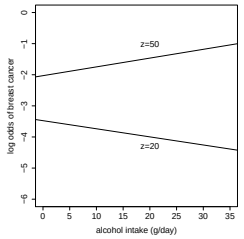
Solution, cont'd

$$\text{logit}(p_{y|x,z}) = \alpha + \beta x + \gamma z + \psi xz$$

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.4406472  0.6923502  -6.414 1.42e-10 ***
alc         -0.0622077  0.0416794  -1.493   0.1356
age          0.0482226  0.0106324   4.535 5.75e-06 ***
alc:age      0.0018085  0.0007812   2.315   0.0206 *
```

$$\text{logit}(p_{y|x,z=20}) = \hat{\alpha} + \hat{\beta} x + \hat{\gamma} \times 20 + \hat{\psi} x \times 20 = -3.48 - 0.026x$$

$$\text{logit}(p_{y|x,z=50}) = \hat{\alpha} + \hat{\beta} x + \hat{\gamma} \times 50 + \hat{\psi} x \times 50 = -2.03 + 0.028x$$



- *Is the difference in slopes relevant?*

Summary

- The logistic regression model is a suitable model for binary outcomes
- Somewhat more difficult to interpret than linear regression...
- ...but restricts the risk of the outcome to the range $(0, 1)$
- Like linear regression, logistic regression
  - can be applied to a wide range of scenarios
  - can be modified and extended in a number of ways
- The logistic regression model is a useful tool to adjust for confounding