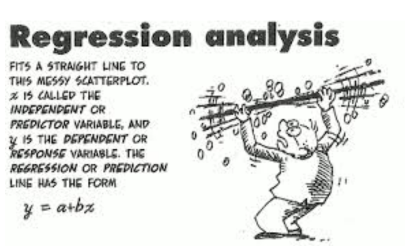
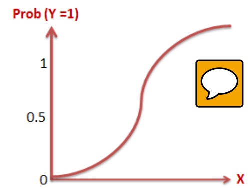
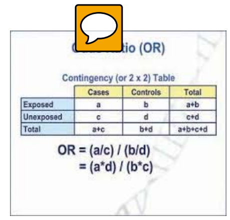
Logistic Regression

Introduction
HL Chapter 1







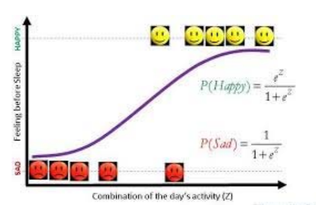
Contingency (or 2 x 2) Table

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

OR = (a/c) / (b/d)
= (a*d) / (b*c)

2×2 tables

Logistic regression



L◉GIT

Linear portion of the logistic regression equation

$$z = \beta_0 + \beta_1 x$$

CAUTION
Risk
Factor

Coefficients

Are they **SIGNIFICANT** ?

Goal

- Build a model to predict or explain a dichotomous outcome (0/1) such as disease or mortality status
- Note: Study subjects with a particular constellation of risk factors may have opposite outcomes
- This is different from linear regression (continuous outcome) where "similar" outcome values are possible

Disease probability

- $\pi(x)$ = probability of the disease given a particular risk factor
- Example, smoking and lung cancer:
  $\pi(smoking = yes)$
  = probability of lung cancer among smokers

Disease probability

- $0 \le \pi(x) \le 1$ (S shaped)
  ➔ linear regression won't work
- Note: Many different regression models could be used to estimate $\pi(x)$

The logistic regression model

- The logistic regression model is a good choice because
  - It can be used easily with available software
  - It provides odds ratios adjusted for confounding
- The logistic regression model is defined as
  $$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

The logit transformation

- The logistic regression model is non-linear
- A transformation called the **logit transformation** can be used to obtain a linear equation

The logit transformation

Intercept

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = ln\left[\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1}{1+e^{\beta_0+\beta_1 x}}}\right] = ln(e^{\beta_0+\beta_1 x}) = \beta_0 + \beta_1 x$$

Risk factor

Slope

Model coefficients

- $\beta_1 = 0$: $\pi(x)$ does not depend on x

Example for lung cancer

- $\beta_1 = 0$: $\pi(eye\ color)$ does not depend on eye color
- I.e. the probability of lung cancer does not depend on eye color

Model coefficients

- $\beta_1 \neq 0$: $\pi(x)$ depends on x

Example for lung cancer

- $\beta_1 \neq 0$: $\pi(smoking)$ depends on smoking
- I.e. the probability of lung cancer depends on smoking

Stat. significance of model coefficients

- If β_1 is statistically significant, the model with *x* predicts the outcome "better" than the model without *x*
- Warning: This does not necessarily mean that the model with *x* predicts the outcome well

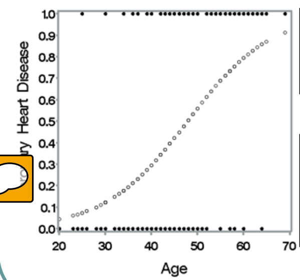
Estimating the model coefficients

- Maximum likelihood estimation
  - Create likelihood function expressing the probability of actually observing the data we observed
  - Choose coefficients β_0 and β_1 such that the likelihood function is maximized
  - Given $\hat{\beta}_0$ and $\hat{\beta}_1$, the outcome predicted by the model mirrors the observed outcome most closely

Example

- Maximum likelihood estimation
  - Find an equation that expresses the probability of observing the lung cancer and smoking data we observed
  - Choose coefficients β_0 and β_1 such that this probability is maximized
  - Given $\hat{\beta}_0$ and $\hat{\beta}_1$, the outcome (lung cancer yes/no) predicted by the model (given smoking status) mirrors the observed data most closely

Example: Coronary Heart Disease (CHD) and age



Dots = observed data

The higher the age, the more dots are at CHD=1

Circles = $\hat{\pi}$(age)
The higher the age, the greater the proportion with CHD=1

Significance tests – Wald test

- Need
  - $\hat{\beta}_1$ = estimated coefficient for variable *x*
  - $\mathrm{SE}(\hat{\beta}_1)$ = estimate of its standard error
- Test statistic: $W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}\,(\hat{\beta}_1)}$
- H_0: $\beta_1 = 0$
- If H_0 is true, then W^2 is χ^2 distributed with 1 df

Example: chd and age

```
libname sdat 'C:\ERHS642';

data chdage; set sdat.chdage; run;

proc logistic descending data=chdage;
  model chd=age;
run;
```

Example: chd and age

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.3095	1.1337	21.9350	<.0001
age	1	0.1109	0.0241	21.2541	<.0001

$p < 0.05$

➔ the model with age predicts chd better than the model without age

Odds ratio and confidence interval

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.117	1.066	1.171

CI does not include 1

➔ The model with age predicts chd better than the model without age

Significance tests – Likelihood ratio test

- Compares
  - "difference" between the model with the variable of interest and the "perfect model"

  to the
  - "difference" between the model without the variable of interest and the "perfect model"

Significance tests – Likelihood ratio test

- Test statistic:

$$G = -2ln\left[\frac{likelihood\ WITHOUT\ x}{likelihood\ WITH\ x}\right]$$

$$= -2[\ln(likelihood\ WITHOUT\ x) - \ln(likelihood\ WITH\ x)]$$

$$= \mathbf{2[\ln(likelihood\ WITH\ x)] - 2[\ln(likelihood\ WITHOUT\ x)]}$$

- H_0: $\beta_1 = 0$
- If H_0 is true, then $\cancel{W^2}$ is χ^2 distributed with 1 df

Example: chd and age

- From SAS: G = 136.663 – 107.353 = 29.31

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	138.663	111.353
SC	141.268	116.563
-2 Log L	**136.663**	**107.353**

Without *x* With *x*

Example: chd and age

Simpler alternative

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	**29.3099**	**1**	**<.0001**
Score	26.3989	1	<.0001
Wald	21.2541	1	<.0001

$p < 0.05$

→ The model with age predicts chd better than the model without age

Wald test vs. likelihood ratio test

- Wald test p-values can be artificially low
- Wald test 95% CIs can be artificially narrow
- This is most likely to occur for small sample sizes but can also happen for larger sample sizes
- Likelihood ratio test is preferable

Wald test vs. likelihood ratio test

- SAS can be used to calculate likelihood ratio-based CIs

```
proc logistic descending data=chdage;
  model chd=age/clodds=both;
run;
```

Wald test vs. likelihood ratio test

- In the chd and age example, the Wald and likelihood-ratio (LR) based CIs are similar

Odds Ratio Estimates for age				
Test	Unit	Estimate	95% Confidence Limits	
Wald	1.0000	1.117	1.066	1.171
LR-based	1.0000	1.117	1.069	1.176

- Unless statistical significance is “borderline-ish”, conclusions are generally the same based on both CIs

Wald test vs. likelihood ratio test

- An asymmetry index can be calculated to estimate whether the two CIs will be much different
- We won’t bother

Assumptions for all tests

- “Large” sample size
- “Reasonable” number of subjects with and without the outcome (e.g., death or disease)
- More on sample size later in the semester