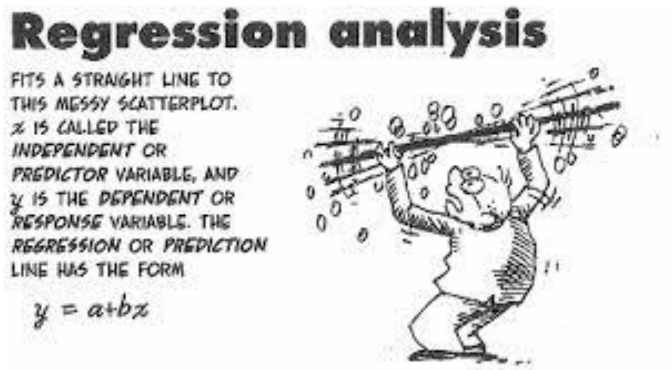
Multiple Logistic Regression

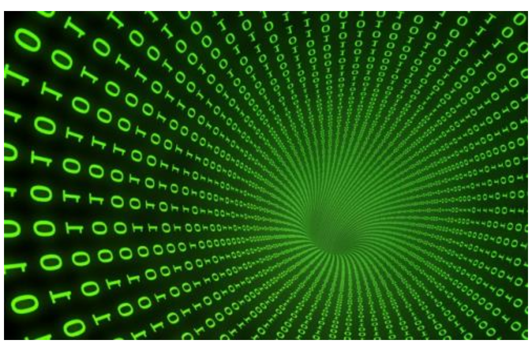
HL Chapter 2

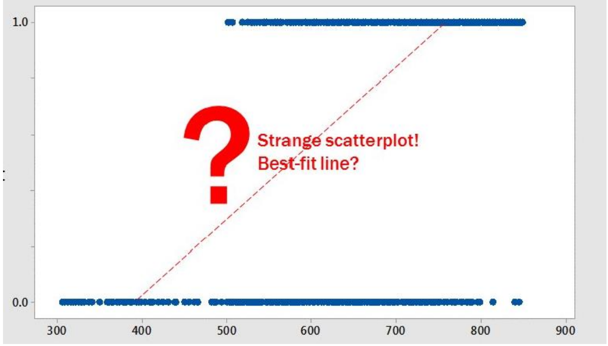

Risk
Factor

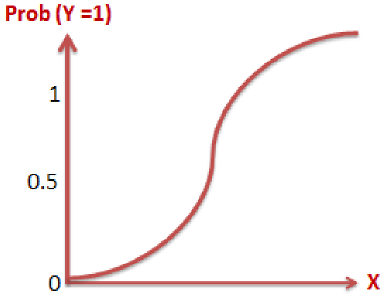

Disease
Health

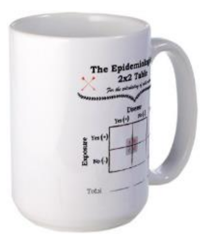

DENOTED
VALUES
ASSUMPTION
LINEAR
DIAGNOSTIC
PREDICT
POPULATION
CAUSALITY
METHOD
ESTIMATION
OBSERVATION
ERROR
PARAMETRIC
REGRESSION
TESTABLE
CAUSAL
SPATIAL
EXPERIMENTER
INFORMATION
ANALYSIS
SET
UNKNOWN
NUMBER
INDEPENDENT
ORDINAL
INFERENCE
MULTIVARIATE
VECTOR
TECHNIQUE
HYPOTHESIS
BINOMIAL
MEASUREMENT
DEPENDENT
ASSUMPTIONS
METHODS
ASYMPTOTIC
PACKAGES
ASSUMED

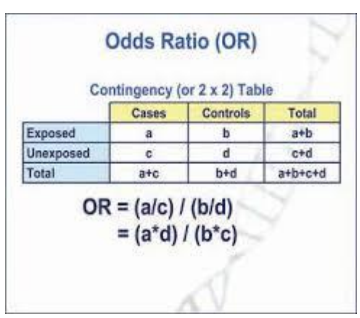

Regression analysis
FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM
y = a+bx




1.0
0.0
Strange scatterplot!
Best-fit line?
300
400
500
600
700
800
900


Prob (Y =1)
1
0.5
0
X





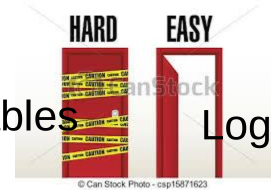
Odds Ratio (OR)

Contingency (or 2 x 2) Table

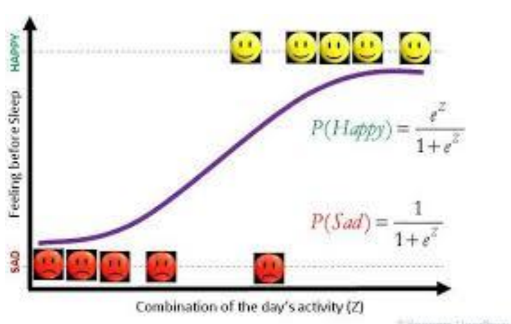
	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

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

2×2 tables



Logistic regression



LOGIT

Linear portion of the logistic regression equation

CAUTION
Risk
Factor

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

Coefficients

Are they SIGNIFICANT ?

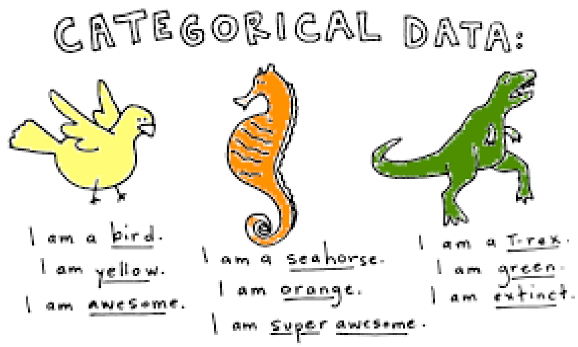
$$z = \beta_0 + \beta_1 x$$

Wait a minute…only one risk factor?

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Coefficients

Are they SIGNIFICANT?



Design (dummy) variables

Race		R2	R3
White (1)		0	0
Black (2)	→	1	0
Other (3)		1	1

Goal

- Build a model to predict or explain a dichotomous outcome (0/1) such as disease or mortality status
- Use more than one predictor

The multivariate model

- The multivariate logistic regression model is defined as

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}$$

(p=number of variables in the model)

The logit transformation

$$g(\tilde{x}) = g(x_1, x_2, \dots, x_p) = \ln\left(\frac{\pi(\tilde{x})}{1-\pi(\tilde{x})}\right) =$$

$$\ln\left[\frac{\dfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_p x_p}}}{\dfrac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_p x_p}}}\right]$$

$$= \ln(e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_p x_p}) = \underbrace{\beta_0}_{\text{Intercept coefficient}} + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Slope coefficients

Design variables (aka dummy variables)

- In SAS, categorical variables are treated as if they were continuous
- Example: GLOW500 data set

RATERISK = self reported risk of fracture

1 = less than others of the same age
2 = same as others of the same age
3 = greater than others of the same age

Design variables (aka dummy variables)

- The variable RATERISK is treated in SAS as if the numbers 1, 2 and 3 were values of a continuous variable such as age
- This doesn't make sense and design variables must be created

Design variable creation

- Most common approach:
  - Select a reference category (e.g., RATERISK = 1) to which the other categories are compared
- If a variable has *c* categories, we need *c-1* design variables

RATERISK	R2	R3
1 = less than others of the same age	0	0
2 = same as others of the same age	1	0
3 = greater than others of the same age	0	1

Design variables

- Example:

ID	RATERISK	R2	R3
1	1 = less than others of the same age	0	0
2	2 = same as others of the same age	1	0
3	3 = greater than others of the same age	0	1
4	2 = same as others of the same age	1	0
5	1 = less than others of the same age	0	0
6	2 = same as others of the same age	1	0
Etc.			

Design variables in SAS
Option 1: Create your own

```
data glow500; set sdat.glow500;
         if raterisk=1 then do; r2=0; r3=0; end;
    else if raterisk=2 then do; r2=1; r3=0; end;
    else if raterisk=3 then do; r2=0; r3=1; end;
run;

proc logistic descending data=glow500;
  model fracture=r2 r3;
run;
```

Design variables in SAS
Option 2: Let SAS create them

```
proc logistic descending data=glow500;

  class raterisk/param=ref ref=first;

  model fracture=raterisk;
run;
```

Results Option 1

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6023	0.2071	59.8311	<.0001
r2	1	0.5462	0.2664	4.2028	0.0404
r3	1	0.9091	0.2711	11.2418	0.0008

Results Option 2

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6023	0.2071	59.8311	<.0001
RATERISK	2	1	0.5462	0.2664	4.2028	0.0404
RATERISK	3	1	0.9091	0.2711	11.2418	0.0008

Stat. significance of design variables

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6023	0.2071	59.8311	<.0001
r2	1	0.5462	0.2664	4.2028	0.0404
r3	1	0.9091	0.2711	11.2418	0.0008

- R2 is only borderline stat. significant at the 0.05 level
- What happens if we exclude R2 from the model?

Stat. significance of design variables

- R3=1 for RATERISK=3
- R3=0 for RATERISK=1 <u>and</u> 2
- R3 compares RATERISK=3 to RATERISK=1 <u>and</u> 2

- Before excluding R2, ee must ensure that this comparison makes sense

RATERISK	R2	R3
1 = less than others of the same age	0	0
2 = same as others of the same age	1	0
3 = greater than others of the same age	0	1

Entering design variables in a model

- Excluding part of a set of design variables may result in combinations of categories that don't make sense (e.g., combining "agree" with "disagree")
- In general, all design variables in a set should be kept in the model even if some are statistically non-significant
- When only part of a set of design variables is included, it is important to determine whether the resulting comparisons make sense

Significance in multivariate models

- SAS uses the Wald test
- Likelihood ratio test can be obtained but it is not automatically provided by SAS and requires calculations

proc logistic descending data=glow500;
  class raterisk/param=ref ref=first;
  model fracture=age weight priorfrac premeno raterisk;
run;

Significance in multivariate models

- WEIGHT and PREMENO are stat. non-significant

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.6057	1.2207	21.0897	<.0001
AGE		1	0.0501	0.0134	13.9660	0.0002
WEIGHT		**1**	**0.00408**	**0.00693**	**0.3470**	**0.5558**
PRIORFRAC		1	0.6795	0.2424	7.8581	0.0051
PREMENO		**1**	**0.1870**	**0.2767**	**0.4565**	**0.4993**
RATERISK	2	1	0.5345	0.2759	3.7539	0.0527
RATERISK	3	1	0.8741	0.2892	9.1381	0.0025

Remove one variable at a time

- Remove WEIGHT

```
proc logistic descending data=glow500;
  class raterisk/param=ref ref=first;
  model fracture=age priorfrac premeno raterisk;
run;
```

Remove one variable at a time

- WEIGHT removed; PREMENO is still non-significant

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.1500	0.9361	30.2653	<.0001
AGE		1	0.0478	0.0128	13.9796	0.0002
PRIORFRAC		1	0.6921	0.2414	8.2238	0.0041
PREMENO		**1**	**0.1926**	**0.2765**	**0.4852**	**0.4861**
RATERISK	2	1	0.5336	0.2759	3.7402	0.0531
RATERISK	3	1	0.8547	0.2868	8.8777	0.0029

Remove one variable at a time

- Remove PREMENO

```
proc logistic descending data=glow500;
  class raterisk/param=ref ref=first;
  model fracture=age priorfrac raterisk;
run;
```

Remove one variable at a time

- Remove PREMENO; the remaining variables are statistically significant at the 0.05 level

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.9905	0.9027	30.5641	<.0001
AGE		1	0.0459	0.0124	13.6179	0.0002
PRIORFRAC		1	0.7002	0.2412	8.4308	0.0037
RATERISK	2	1	0.5485	0.2750	3.9786	0.0461
RATERISK	3	1	0.8657	0.2862	9.1500	0.0025

Removing >1 variable at a time: Likelihood ratio (LR) test

- Test statistic:

$$G = -2ln\left[\frac{likelihood\ AFTER\ removing\ variables}{likelihood\ BEFORE\ removing\ variables}\right]$$

$$= -2[\ln(likelihood\ AFTER\ removing\ variables)$$

Removing >1 variable at a time: Likelihood ratio (LR) test

- H_0: Coefficients of variables of interest equal to 0
- If H_0 is true, then G is χ^2 distributed with df = number of removed variables

Removing >1 variable at a time: Likelihood ratio (LR) test

- Model with all 6 variables including WEIGHT and PREMENO

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	44.2598	6	<.0001
Score	44.5299	6	<.0001
Wald	40.4058	6	<.0001

p<0.05
➔ Model with 6 variables is significantly better than model without any variables

Removing >1 variable at a time: Likelihood ratio (LR) test

- Model with 4 variables (WEIGHT and PREMENO removed)

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	43.4363	4	<.0001
Score	43.9332	4	<.0001
Wald	39.8785	4	<.0001

p<0.05
➔ Model with 4 variables is significantly better than model without any variables

Removing >1 variable at a time: Likelihood ratio (LR) test

- Is the model with 6 variables significantly better than the model with 4 variables?

Test	Chi-Square	DF
Likelihood Ratio	**44.2598**	**6**

$$2\ln\left(\begin{matrix} likelihood\ before \\ removing\ variables \end{matrix}\right)$$

Test	Chi-Square	DF
Likelihood Ratio	**43.4363**	**4**

$$2\ln\left(\begin{matrix} likelihood\ after \\ removing\ variables \end{matrix}\right)$$

- Likelihood-ratio test:

  G = 44.2598 – 43.4363 = 0.8235

- df = 2 (2 variables removed)

Removing >1 variable at a time: Likelihood ratio (LR) test

```
data pval;
  p=1-probchi(0.8235,2);
run;

proc print data=pval; run;
```

G df

p = 0.6625 ➔ The 6 variable model is not significantly better than the 4 variable model

- Note: Wald or Score test could have been used instead

IMPORTANT: Missing values

- When a variable with missing values is removed from the model, the number of observations differs between the two models being compared
- To correctly compare two models, they must be based on the same set of observations
- See in-class assignment 3 for an example