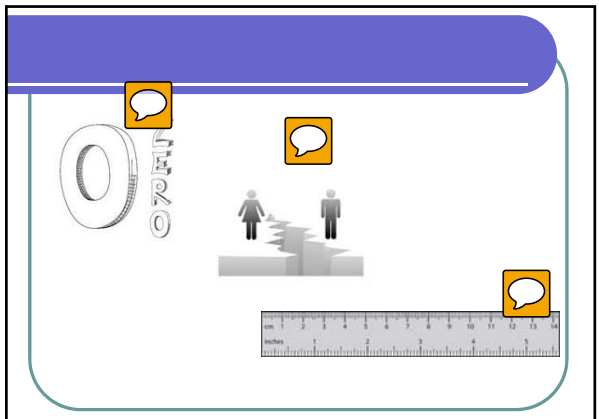
Numerical problems

HL Chapter 4 – part 2



I. Zero cells

Example

- Assume we are working with a data set based on which we get the following results from proc freq

Table of race by D			
race	D		
	0	1	Total
1	100	20	120
2	10	1	11
3	50	3	53
4	2	0	2
Total	162	24	186

Nobody of race "4" has the outcome

- How does this affect the logistic regression analysis?

Zero cells – Effect on logistic regression analysis

- Log Window:

WARNING: There is possibly a quasi-complete separation of data points. The maximum likelihood estimate may not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Zero cells – Effect on logistic regression analysis

- Results Window:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6094	0.2449	43.1715	<.0001
race	2	1	-0.6931	1.0770	0.4142	0.5199
race	3	1	-1.2040	0.6429	3.5070	0.0611
race	4	1	-12.5212	827.8	0.0002	0.9879

Huge coefficient and std error

Zero cells – Effect on logistic regression analysis

- Results window

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
race 2 vs 1	0.500	0.061	4.128
race 3 vs 1	0.300	0.085	1.058
race 4 vs 1	<0.001	<0.001	>999.999

Useless OR and 95% CI

- Zero cell ➔ Model falls apart
- Collapse categories or eliminate the category with the zero cell
- Adding ½ to each cell is generally not recommended

Possible reasons for zero cells

- Random error (sample size too small)
- Systematic error (accidentally excluded subjects in a certain category)
- True absence of subjects in the category

II. Complete separation

- Happens when one or more covariates perfectly predict the outcome resulting in zero cells
- In most cases this is not due to the fact that the covariate is a perfect predictor
- It is usually due to random or systematic error introduced during data collection or due to overfitting the model

Complete separation - Example

Example: In a study population,
- All breast cancer cases are female
- All controls are males

	Breast cancer	No breast cancer	Total
Female	100	0	100
Male	0	100	100
Total	100	100	200

Complete separation – Effect on logistic regression analysis

- Log window

WARNING: There is a complete separation of data points. The maximum likelihood estimate does not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Complete separation – Effect on logistic regression analysis

- Results window

Model Convergence Status
Complete separation of data points detected.

Warning: The maximum likelihood estimate does not exist.

Warning: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Complete separation – Effect on logistic regression analysis

- Results window

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	9.2027	9.9631	0.8532	0.3556
sex	1	-18.4055	14.0899	1.7064	0.1915

huge

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	<0.001	<0.001	>999.999

useless

Complete separation – Effect on logistic regression analysis

- Complete separation ➔ zero cells ➔ model falls apart
- The variable cannot be included in the model

III. Quasi-complete separation

- Happens when one or more covariates <u>almost</u> perfectly predict the outcome

Quasi-complete separation - Example

Example: In a study population,

- Almost all breast cancer cases are female
- Almost all controls are males

	Breast cancer	No breast cancer	Total
Female	97	3	100
Male	3	97	100
Total	100	100	200

Quasi-complete separation – Effect on logistic regression analysis

- Log window

No warnings

- Results window

No warnings

Quasi-complete separation – Effect on logistic regression analysis

- Results window

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.4760	0.5862	35.1636	<.0001
sex	1	-6.9520	0.8290	70.3272	<.0001

↖ very large

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	<0.001	<0.001	0.005

useless

IV. Collinearity between independent variables

- Happens when two or more variables are identical

Collinearity - Example

Example:

- In a study population, weight is recorded in kilograms and pounds
- Assume the outcome is disease D
- Further assume that both weight variables are included in the logistic regression model

Collinearity – Effect on logistic regression analysis

- Log window

No warnings

Collinearity – Effect on logistic regression analysis

- Results window

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2620	2.6917	0.2198	0.6392
weight_kg	1	-0.0212	0.0451	0.2210	0.6383
weight_lb	0	0	.	.	.

Collinearity – Effect on logistic regression analysis

- Results window

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
weight_kg	0.979	0.896	1.070

- Only one of the two collinear variables is included
- Remove the other collinear variable from the SAS model statement