2.2.5 Pitfalls of Regression Modelling

Five issues to bear in mind in ANOVA, Regression and General Linear Modelling.

1. Model assumptions
2. Data transformations
3. Model selection
4. Multicollinearity
5. Predicting beyond the range of the covariates

See Handout.

Non-Parametric Statistics

All of the previous statistical analysis methods studied (t-tests, ANOVA, Regression, General Linear Modelling) have depended heavily on **distributional assumptions**.

i.e. we assume that the data are **Normally distributed**.

We now seek statistical procedures that do not rely on this strong assumption. We term these methods

NON-PARAMETRIC

or

DISTRIBUTION-FREE

They substitute **large sample approximations** for the distributional assumptions.

3.1 Distribution-free tests for Categorical Data

Categorical data are data in which experimental units are allocated to one of a number of categories according to their characteristics. The categories are defined by one or more factors

Examples:

- Female/Male - two categories
- Smoker/Former Smoker/Non Smoker - three categories.

Doll and Hill Data

Table 13.11. Smokers and non-smokers among male cancer patients and controls (Doll and Hill 1950)

	Smokers	Non-smokers	Total
Lung cancer	647	2	649
Controls	622	27	649

Juvenile Delinquency and Spectacle-Wearing

Table 10.14 Spectacle wearing among juvenile delinquen and non-delinquents who failed a vision test (Weindling *al* 1986)

		Juvenile delinquents	Non delinquents	Total
Spectacle wearers	Yes	1	5	6
	No	8	2	10
	Tota	9	7	16