Overdispersion

- Recall 1: if the binomial GLM model specification is correct, i.e.,
  \[ y_x \sim B(n_x, p_x = g^{-1}(\eta_x)) \]
  \[ \Rightarrow E(y_x) = n_x p_x \quad \text{and} \quad \text{Var}(y_x) = n_x p_x (1 - p_x) \leq \frac{n_x}{p_x} \quad \text{max at} \quad p_x \leq \frac{1}{2} \]
  \[ \Rightarrow D (\text{deviance}) \sim \chi^2_{k-p} \Rightarrow \text{can perform goodness-of-fit test} \]

- \[ Q: \text{why no such goodness-of-fit test for Normal linear model if no further assumption/information about } \sigma^2 \text{ is offered?} \]

Recall 2: \[ D \approx \chi^2 = \sum (r_i^2)^2 \]

- \[ Q: \text{what cause large } D? \Rightarrow \text{may suspect } \text{Var}(y_x) \gg n_x p_x (1 - p_x) \]

- Some possible explanation for large deviance (c.f., possible reasons causing \( \hat{\sigma}^2 \) larger than \( \sigma^2 \) in Normal linear model)

  - Sparse data (i.e., \( n_x \)'s too small) \( \Rightarrow D \sim \chi^2_{k-p} \) is questionable
  - Presence of outliers (can be detected in diagnostics for GLM)

- For larger number of outliers, we might conclude that they are unexceptional \( \Rightarrow \text{something amiss with other structures} \)

Wrong \( X/\beta \) structure

- Predictors not transformed/combined in a correct way (diagnostics for GLM (future lecture) can help)

- Important predictors not included (\( \Rightarrow \text{cause overdispersion if these important predictors are not available} \)

- Deficiencies in the random part of the model

[Note: \( y_x \sim B(n_x, p_x) \) when the corresponding \( n_x \) \( z_x \)'s are independent and identically distributed as \( B(1, p_x) \)]

overdispersion: \( \text{Var}(y_x) \gg n_x p_x (1 - p_x) \)

underdispersion: \( \text{Var}(y_x) \ll n_x p_x (1 - p_x) \) (in rare cases)

- Violation of constant \( p_x \) assumption

- Example: in shuttle disaster case, position of the O-ring on the booster rocket may have some effect on the failure probability. Yet, this variable was not recorded.
**Q:** Why can the heterogeneity cause overdispersion?

Consider the situation: a population is divided into clusters, sample size = \( m \), cluster size = \( k \), number of clusters = \( l = m/k \).

- number of successes in cluster \( i \): \( s_i \sim \text{B}(k, p_i) \)
- \( p_i \): random variable with mean \( p \) and variance \( \tau^2 p(1-p) \)

\[
y = s_1 + \ldots + s_l \Rightarrow E(y) = mp
\]

But, \( \text{Var}(y) = \sum \text{Var}(s_i) = \sum \{E(\text{Var}(s_i|p_i)) + \text{Var}(E(s_i|p_i))\} \)

\[
= [1+(k-1)\tau^2]mp(1-p) > mp(1-p)
\]

- Note: for \( m=1 \) (sparse case), this problem cannot arise

**Violation of dependence assumption**

- Causing overdispersion: For example, response has a common cause, say a disease is influenced by genes, the responses will tend to be positively correlated

- Causing underdispersion: For example, when food supply is limited, probability of survival of an animal may be increased by the death of others

**Q:** how to model overdispersion and do analysis?

- Introduce an additional dispersion parameter \( \sigma^2 \), i.e.,

\[
\text{Var}(y) = \sigma^2 np(1-p) \quad \text{notice the similarity to linear model (standard binomial case } \Rightarrow \sigma^2=1; \text{ overdispersion } \Rightarrow \sigma^2>1)\]

- \( \sigma^2 \) may be estimated using \( \hat{\sigma}^2 = \chi^2 (X_1 - p) \) (an alternative is to use the deviance in place of the \( \chi^2 \), which is not very recommended as it may not be consistent)

- Estimation of \( \beta \) is unaffected since \( E(y) \) is not changed

- But, \( \text{Var}(\hat{\beta}) = \sigma^2 (X'W) X^{-1} \)

- Because now difference in deviances \( \sim \sigma^2 \chi^2 \), to compare models, i.e., test \( H_0: S \text{ v.s. } H_1: L \mid S \):

\[
F = \frac{(D_S - D_L)/(df_S - df_L)}{\hat{\sigma}^2} \sim F_{df_S - df_L, k-p}
\]

- No goodness-of-fit test is possible

- This dispersion parameter method is appropriate when the covariate classes are roughly equal in size (\( n_1 \approx n_2 \approx \ldots \approx n_k \)); otherwise, more sophisticated methods should be used, such as beta-binomial method, quasi-likelihood, …

**Reading:** F, 2.11
Matched Case-Control Studies

- **Q:** In a case-control study, how should we choose the controls if there exist some confounding variables, say age and sex, that may affect the outcome in addition to the risk factors?

  - **Approach 1:** record and regard confounding variables as risk factors in the logistic regression analysis (however, we may not be interested in the effect of the confounding variables)

  - **Approach 2:** confounding variables are explicitly adjusted for in the design

### Matched case-control design (MCCD)

Matched case-control design (MCCD): match each case with one or more controls that have the same or similar values of some set of potential confounding variables. A group of a case and its corresponding controls is called a *matched set*. For example,

<table>
<thead>
<tr>
<th>1st case</th>
<th>2nd case</th>
<th>3rd case</th>
</tr>
</thead>
<tbody>
<tr>
<td>age=20; sex=male</td>
<td>age=20; sex=female</td>
<td>age=70; sex=female</td>
</tr>
<tr>
<td>( D^c )</td>
<td>( D^c )</td>
<td>( D^c )</td>
</tr>
<tr>
<td>( X^c )</td>
<td>( X^c )</td>
<td>( X^c )</td>
</tr>
<tr>
<td>( y_{100} )</td>
<td>( y_{200} )</td>
<td>( y_{200} )</td>
</tr>
<tr>
<td>( X )</td>
<td>( X )</td>
<td>( X )</td>
</tr>
<tr>
<td>( y_{110} )</td>
<td>( y_{200} )</td>
<td>( y_{200} )</td>
</tr>
</tbody>
</table>

#### 1: \( M \) MCCD: \( M \) controls for each case

- \( M \) typically small, can vary in size in every matched set
- Each additional control yields a diminished return in terms of increased efficiency in estimating risk factors
- It is usually not worth exceeding \( M = 5 \)

#### Some disadvantages of matched case-control design

- Lose the possibility of discovering the effects of the confounding variable
  \( \Rightarrow \) cannot estimate
- The data will likely be far from a random sample of the population of interest

#### Analysis of MCCD

- For individual \( i \) in the \( j \)th matched set, observe the value of risk factor \( x_{ij} \)
- Consider a 1:\( M \) MCCD with \( n \) matched sets, denote
  \( i = 0 \Rightarrow \) case and \( i = 1, \ldots, M \Rightarrow \) control