Case Control Analysis Summary Sheet

1 Introduction

- While experimental studies like randomized clinical trials are considered the gold standard for demonstrating causality, they are not suited for demonstrating evidence in favor or against every scientific question
- Case-control studies are observational studies that select cases (disease) and then controls separately to try and identify the effect of some exposure on a patient's health status retrospectively
- This is possible since the exposure odds ratio is equivalent to the odds ratio (see binomial data sheet)
- These studies are popular amongst investigations into *rare* diseases

1.1 Setup

- select n_1 diseased, n_0 non-diseased
- Objective of analysis: Contrast $\vec{x}_i$ beteween cases $Y_i = 1$ and controls $Y_i = 0$

  adjust for sampling fractions $\tau_0 \equiv P(S_i = 1|Y_i = 0), \tau_1 \equiv P(S_i = 1|Y_i = 1)$
- Model:

$$\log(\frac{p_i}{1-p_i}) = x_i^T\beta \iff p_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

1.2 Maximum Likelihood Estimation

$$L_i(\vec{\beta}) \propto P(x_i|Y_i = 1, S_i = 1)^{Y_i} P(x_i|Y_i = 0, S_i = 1)^{1-Y_i}$$

- Conveniently, this estimation is equivalent to the prospective/retrospective cohort logistic regression model[1]
- Important to note that β_0^* the intercept derived in retrospective sampling logistic regression needs to be adjusted for by the sampling fraction

$$\begin{aligned} \log(\text{odds}(\vec{x}_i)) &= \vec{x}_i'\beta^* \\ &= \vec{x}_i'\beta + \log(\frac{\tau_0}{\tau_1}) \end{aligned}$$

[1] Prentice and Pyke (1979)

1.3 Example: Lung Cancer Case Control Study

Set-up:
In order to understand how different study designs and analysis estimate the truth, assume the following to be true for the next example:

$RR = 2$ $\qquad$ $P(X=1) = 0.2$

$P(Y=1|X=0) = \pi_0 = 0.05$ $\qquad$ $P(Y=1|X=1) = \pi_1 = 0.1$

Model:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$$

$\beta_0 = \log\{\frac{.05}{.95}\} = -2.944$ $\qquad$ $\beta_0 + \beta_1 = \log(\frac{.1}{.9})$

$\beta_1 = .747$ $\qquad$ $\exp\{\beta_1\} = 2.11$

Case-Control Data

		Health Status		
		$Y=0$	$Y=1$	total
Treatment	$X=0$	4022	3358	7380
	$X=1$	978	1642	2620
	total	5000	5000	10000

- $\hat{\pi}_0 = \frac{3358}{7380} \approx 0.4555$
- $\hat{\pi}_1 = \frac{1642}{2620} \approx 0.627$
- $\hat{\pi}^* = P(Y=1) = \frac{5000}{10000} = 0.5$
- $\hat{RD} = 0.627 - 0.455 = 0.172$
- $\hat{RR} = \frac{0.627}{0.455} = 1.378$
- $\hat{OR} = \frac{4022 \times 1642}{3358 \times 978} \approx 2.011$
- $\hat{\beta}_0 = \ln\left(\frac{3358/7380}{(4022/7380)}\right) \approx -0.178$
- $\hat{\beta}_0 + \hat{\beta}_1 = \ln\left(\frac{1642/2620}{(978/2620)}\right) \approx 0.5181$
- $\hat{\beta}_1 = .6961;$ $\qquad$ $\exp\{\hat{\beta}_1\} = 2.01$
- Prevalence: $P(Y=1) = 0.06$

  see work below
- Sampling fractions ratio $\frac{\tau_1}{\tau_0} = \frac{.94}{0.06} \approx 15.67$
- $\hat{\beta}_0 - \log(\frac{\tau_1}{\tau_0}) = -.178 - \underbrace{\ln(15.67)}_{2.75} = -2.928$

  Note the more accurate estimate

Prevalence Derivation

$$\begin{aligned}
P(Y=1) &= P(Y=1|X=0)P(X=0) + P(Y=1|X=1)P(X=1) \\
&= P(Y=1, X=0) + P(Y=1, X=1) \\
&= 0.05 \times .8 + .1 \times .2 \\
&= 0.06
\end{aligned}$$

Sampling Fraction Derivation

$$\begin{aligned}
\tau_1 = P(S=1|Y=1) &= \frac{P(Y=1|S=1)P(S=1)}{\times} PY=1) \\
&= \frac{0.5 \times \frac{10000}{N}}{0.06} \\
\tau_0 = P(S=1|Y=0) &= \frac{P(Y=0|S=0) \times P(Y=0)}{P(Y=0)} \\
&= \frac{.5 \times \frac{10000}{N}}{.94} \\
\Rightarrow \frac{\tau_1}{\tau_0} &= \frac{0.5 \times \frac{10000}{N}}{0.06} \times \frac{.94}{.5 \times \frac{10000}{N}} \\
&= \frac{.94}{0.06} = 15.6\bar{6}
\end{aligned}$$