020 - Logistic Regression I

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on November 17, 2020

Confounding revisited

- The study on success of different kidney stone removal procedures reported by Charig et al. (1986) provides a nice and clean example of confounding, where the direction of the estimated effect size is reversed by introducing stratification.
- The procedure, either open surgery, percutaneous nephrolithotomy (PN, a keyhole surgery procedure) or extracorporeal shock wave lithotripsy (ESWL), was defined to be successful if stones were eliminated or reduced to less than 2 mm after three months.
- The study collected cases of kidney stones treated at a particular UK hospital during $1972 - 1985$.
- 350 of these cases were treated with open surgery and 350 with percutaneous nephrolithotomy.

Outcome data

- The counts of successes for the two surgical procedures were:

	Unsuccessful	Successful	Total
Open surgery	77	273	350
PN	61	289	350
Total	138	562	700

- Empirical odds ratio for the failure of the procedure is given by

$$\log\left(\frac{77/273}{61/289}\right) = \log\left(\frac{77 \times 289}{61 \times 273}\right) \approx 0.290$$

  and its standard error by

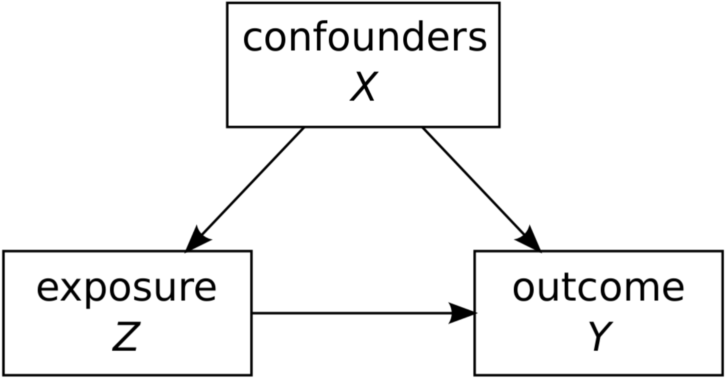
$$\sqrt{\frac{1}{77} + \frac{1}{61} + \frac{1}{273} + \frac{1}{289}} \approx 0.191$$

- Open surgery appears to have higher odds of failure, although the log-odds ratio estimate is smaller than two times the standard error.

Interpretation

- Based on these results, do we have evidence that percutaneous nephrolithotomy is more effective than open surgery?
- It certainly is less invasive.
- Similar pattern can be observed for instance for kidney (and other) cancer surgeries; the patients treated with the less invasive procedure have better outcomes, at least in the short term.
- What is really going on here?
- Remember that the treatment procedure was not randomized; the cases were ascertained from the hospital records.
- Recall 'the triangle'.

The triangle



What are X, Y and Z in the present example?

Outcomes by kidney stone size

- Below are the same outcomes tabulated by the size of the kidney stone (smaller than 2 cm/ at least 2 cm in diameter)

< 2 cm	Unsuccessful	Successful	Total
Open surgery	6	81	87
PN	36	234	270
Total	42	315	357

≥ 2 cm	Unsuccessful	Successful	Total
Open surgery	71	192	263
PN	25	55	80
Total	96	247	343

Stratum-specific odds ratios

- The empirical odds ratio and its standard error in the < 2 cm group are
$$\log\left(\frac{6 \times 234}{36 \times 81}\right) \approx -0.731$$
and
$$\sqrt{\frac{1}{6} + \frac{1}{36} + \frac{1}{81} + \frac{1}{234}} \approx 0.459$$
- Similarly, these numbers in the ≥ 2 cm group are
$$\log\left(\frac{71 \times 55}{25 \times 192}\right) \approx -0.206$$
and
$$\sqrt{\frac{1}{71} + \frac{1}{25} + \frac{1}{192} + \frac{1}{55}} \approx 0.278$$
- Now open surgery appears to have lower odds of failure within the strata.

Interpretation

- Open surgery was much more common in the ≥ 2 cm group, as was the failure of surgery.
- This is not surprising; presumably larger stones are more difficult to remove, whilst also requiring a more invasive procedure.
- Prima facie this seems to be an example of confounding by indication, with kidney stone size being part of the indication for the choice between open and keyhole surgery.
- Was kidney stone size known before the choice of the procedure, or was the indication related to something else, perhaps symptomatic of the size?
- How exactly were the cases selected?

Stratified analysis

- It is obvious that the first pooled analysis was confounded.
- Within stratum estimates are more valid.
- How can we combine the results across strata without re-introducing the confounding?
- We have to take a weighted average of the stratum-specific log-odds ratios.
- Let now $\log \hat{\theta}_0$ and $\log \hat{\theta}_1$ be the empirical log-odds ratios within the < 2 cm and ≥ 2 cm groups, respectively.
- Weighted average of these is given by

$$\log \hat{\theta} = \frac{w_0 \log \hat{\theta}_0 + w_1 \log \hat{\theta}_1}{w_0 + w_1} = \frac{\sum_{j=0}^{1} w_j \log \hat{\theta}_j}{\sum_{j=0}^{1} w_j}$$

- How to choose the weights?

The Woolf 1955 method

- Presumably the choice of the weights must depend on the stratum sizes since a large stratum would be more informative than a small one, requiring larger weight in the combined estimate.
- In fact, the quantity $\frac{1}{S^2}$ is known as the observed information.
- Accordingly, in the Woolf 1955 method for stratified analysis, the weights are chosen as

$$w_j = \frac{1}{\frac{1}{D_{1j}} + \frac{1}{D_{0j}} + \frac{1}{H_{1j}} + \frac{1}{H_{0j}}}$$

- Now we have

$$w_0 = \frac{1}{\frac{1}{6} + \frac{1}{36} + \frac{1}{81} + \frac{1}{234}} \approx 4.738$$

and

$$w_1 = \frac{1}{\frac{1}{71} + \frac{1}{25} + \frac{1}{192} + \frac{1}{55}} \approx 12.907$$

Combined estimate and its standard error

- With these weights, the combined estimate is given by

$$\log \hat{\theta} = \frac{4.738 \times -0.731 + 12.907 \times -0.206}{4.738 + 12.907} \approx -0.347$$

- Standard error is the square root of the inverse of the total information.
- Information is additive.
- Thus, the standard error of the combined estimate is given by

$$\sqrt{\frac{1}{\frac{1}{S_0^2} + \frac{1}{S_1^2}}} = \sqrt{\frac{1}{w_0 + w_1}} = \sqrt{\frac{1}{4.738 + 12.907}} \approx 0.238$$

Limitations of stratified analysis

- In the similar Mantel-Haenszel method, the stratum-specific weights are given by $w_j = \frac{H_{1j}D_{0j}}{N_j}$, where N_j is the total N of table j
- The Woolf and Mantel-Haenszel methods are applicable only when there are no zero cell counts in any of the confounder-conditional 2×2 -tables.
- This becomes an issue whenever there are a large number of confounder strata.
- Usually one would have a large number of potential confounders, some of which are continuous-valued, so cross-stratifying across all of these quickly becomes infeasible.
- The problem needs to be re-parametrized in more parsimonious way.

Another example

- Senn (2003) used these diabetes cohort data to demonstrate *Simpson's paradox*:

	Dead	Censored	Total
Type II	218	326	544
Type I	105	253	323
Total	323	579	902

- Empirical all-cause mortality odds ratio is given by

$$\log\left(\frac{218 \times 253}{105 \times 326}\right) \approx 0.477$$

and its standard error by

$$\sqrt{\frac{1}{218} + \frac{1}{105} + \frac{1}{326} + \frac{1}{253}} \approx 0.145$$

- Type II diabetes patients seem to have higher mortality.

Mortality outcomes by age group

- Below are the same outcomes tabulated by age:

≤ 40	Dead	Censored	Total
Type II	0	15	15
Type I	1	129	130
Total	1	144	145

> 40	Dead	Censored	Total
Type II	218	311	529
Type I	104	124	228
Total	322	435	757

- There is only one death and very few type II diabetes patients in the ≤ 40 age group.
- Obviously, age is a determinant of both the type of diabetes and mortality.

Stratum-specific log-odds ratios

- The empirical log-odds ratio and its standard error in the ≤ 40 group are

$$\log\left(\frac{0 \times 129}{1 \times 15}\right) = -\infty$$

and

$$\sqrt{\frac{1}{0} + \frac{1}{1} + \frac{1}{15} + \frac{1}{129}} = \text{ undefined}$$

- These numbers in the > 40 group are

$$\log\left(\frac{218 \times 124}{104 \times 311}\right) \approx -0.179$$

and

$$\sqrt{\frac{1}{71} + \frac{1}{25} + \frac{1}{192} + \frac{1}{55}} \approx 0.160$$

- Type II diabetes patients over 40 years of age seem to have lower mortality (although not significantly so).

Interpretation

- Now stratified analysis is not feasible; the ≤ 40 table is not informative of the mortality log-odds ratio.
- Given the row totals, the observed data comprise four death counts, with the numbers of censored given by the row total minus death count.
- From four observations, we can estimate at most four parameters; the stratum-specific odds ratios both involve two odds parameters.
- However, now one of the observed counts is zero, so we are reduced to estimating only three parameters.
- The problem needs to be re-parametrized in a more parsimonious way.

Parametrization in terms of risk

- Introduce a variable X to denote the age group, with $X = 0$ for the ≤ 40 group and $X = 1$ for the > 40 group.
- Now we have four risk parameters π_{ZX}, one for each level of Z and X :

$$\begin{array}{lccc} & & \text{Dead} & \text{Censored} \\ X = 0: & Z = 1 & \pi_{10} & 1 - \pi_{10} \\ & Z = 0 & \pi_{00} & 1 - \pi_{00} \end{array}$$

$$\begin{array}{lccc} & & \text{Dead} & \text{Censored} \\ X = 1: & & \pi_{11} & 1 - \pi_{11} \\ & & \pi_{01} & 1 - \pi_{01} \end{array}$$

- Risk parameter is a probability, taking values in the $[0, 1]$-interval.

Parametrization in terms of odds

- Alternatively, we may opt to parametrize in terms of odds, that is, risk divided by one minus itself:

$$
\begin{array}{lccc}
 & & \text{Dead} & \text{Censored} \\
X=0: & Z=1 & \frac{\pi_{10}}{1-\pi_{10}} & \frac{1-\pi_{10}}{\pi_{10}} \\
 & Z=0 & \frac{\pi_{00}}{1-\pi_{00}} & \frac{1-\pi_{00}}{\pi_{00}}
\end{array}
$$

$$
\begin{array}{lccc}
 & & \text{Dead} & \text{Censored} \\
X=1: & Z=1 & \frac{\pi_{11}}{1-\pi_{11}} & \frac{1-\pi_{11}}{\pi_{11}} \\
 & Z=0 & \frac{\pi_{01}}{1-\pi_{01}} & \frac{1-\pi_{01}}{\pi_{01}}
\end{array}
$$

- An odds parameter may take values in the $[0, \infty]$-interval.

Parametrization in terms of log odds

- Or we may prefer log-odds:

$$
\begin{array}{llll}
 & & \text{Dead} & \text{Censored} \\
X=0: & Z=1 & \log\frac{\pi_{10}}{1-\pi_{10}} & \log\frac{1-\pi_{10}}{\pi_{10}} \\
 & Z=0 & \log\frac{\pi_{00}}{1-\pi_{00}} & \log\frac{1-\pi_{00}}{\pi_{00}}
\end{array}
$$

$$
\begin{array}{llll}
 & & \text{Dead} & \text{Censored} \\
X=1: & Z=1 & \log\frac{\pi_{11}}{1-\pi_{11}} & \log\frac{1-\pi_{11}}{\pi_{11}} \\
 & Z=0 & \log\frac{\pi_{01}}{1-\pi_{01}} & \log\frac{1-\pi_{01}}{\pi_{01}}
\end{array}
$$

- A log-odds parameter may take values in the $[-\infty, \infty]$-interval.
- Whichever way, there are still four parameters.
- How to reduce the number of parameters?

Regression

- Clayton & Hills (1993, p. 217)

  *A common theme in all these situations is change from the original parameters to new parameters which are more relevant to the comparisons of interest. This change can be described by the equations which express the old parameters in terms of the new parameters. These equations are referred to as **regression** equations, and the statistical model is called a **regression model**.*

- Now the old parameters are the four log-odds:

$$\log \frac{\pi_{ZX}}{1 - \pi_{ZX}}$$

Regression equation

- Reparametrizing the log-odds is referred to as logistic regression.
- In the ongoing example we may take

$$\log\left(\frac{\pi_{ZX}}{1-\pi_{ZX}}\right) = \alpha + \beta Z + \gamma X$$

- The original four parameters are now expressed in terms of three new parameters: an intercept term α and regression coefficients β and γ.
- The function $\log \frac{\pi}{1-\pi}$ is referred to as the logit transformation of the risk parameter π.
- Thus, the same model can be specified as a reparametrization of the risk parameter together with the *logit link* function:

$$\text{logit}\left(\pi_{ZX}\right) = \alpha + \beta Z + \gamma X$$

Parametrization in terms of regression parameters

- Through the previous model specification we get the log-odds tables

		Dead	Censored
$X = 0:$	$Z = 1$		
	$Z = 0$		

		Dead	Censored
$X = 1:$	$Z = 1$		
	$Z = 0$		

- How to interpret the new parameters?
- Suppose that we are interested in the mortality odds ratio for type II vs. type I diabetes patients, controlling for age.
- In other words, our parameter of interest is

$$\frac{\frac{\pi_{1X}}{1-\pi_{1X}}}{\frac{\pi_{0X}}{1-\pi_{0X}}}$$

Interpretation of the regression coefficient

- Through the regression equation we get

$$
\begin{aligned}
\frac{\frac{\pi_{1X}}{1-\pi_{1X}}}{\frac{\pi_{0X}}{1-\pi_{0X}}} &= \frac{e^{\alpha+\beta+\gamma X}}{e^{\alpha+\gamma X}} \\
&= \frac{e^{\alpha} e^{\beta} e^{\gamma X}}{e^{\alpha} e^{\gamma X}} \\
&= e^{\beta} \\
\Leftrightarrow \log\left(\frac{\frac{\pi_{1X}}{1-\pi_{1X}}}{\frac{\pi_{0X}}{1-\pi_{0X}}}\right) &= \beta
\end{aligned}
$$

- The regression coefficient β is a log-odds ratio.
- Exponentiating it gives the odds ratio of interest.
- Having understood regression as a transformation of the original parameters, where and when does the observed outcome data come into play?

Observed data

- The observed data are the four death counts:

		Dead	Censored	Total
$X=0:$	$Z=1$	D_{10}	$N_{10}-D_{10}$	N_{10}
	$Z=0$	D_{00}	$N_{00}-D_{00}$	N_{00}

		Dead	Censored	
$X=1:$	$Z=1$	D_{11}	$N_{11}-D_{11}$	N_{11}
	$Z=0$	D_{01}	$N_{01}-D_{01}$	N_{01}

- These data entered into `R` are:

```
dead <- c(0,1,218,104)
censored <- c(15,129,311,124)
z <- c(1,0,1,0); x <- c(0,0,1,1)
cbind(dead,censored,z,x)

##      dead censored z x
## [1,]    0       15 1 0
## [2,]    1      129 0 0
## [3,]  218      311 1 1
## [4,]  104      124 0 1
```

Deterministic and stochastic model components

- The regression equation specifies the deterministic part of the model.
- This is defined in terms of parameters, conditional on the values of Z and X.
- To complete the model specification, we need to specify the stochastic component of the model, a statistical distribution for the outcome D_{ZX}.
- It is already obvious that the appropriate distribution is

$$D_{ZX} \sim \text{Binomial}\left(N_{ZX}, \pi_{ZX}\right)$$

- Here the risk π_{ZX} is given by the regression equation as (verify)

$$\pi_{ZX} = \frac{e^{\alpha+\beta Z+\gamma X}}{1+e^{\alpha+\beta Z+\gamma X}} = \frac{1}{1+e^{-(\alpha+\beta Z+\gamma X)}}$$

- This inverse transformation is the so-called *expit* function:

$$\pi_{ZX} = \text{logit}^{-1}(\alpha+\beta Z+\gamma X) = \text{expit}(\alpha+\beta Z+\gamma X)$$

Fitting the model

- We have now specified the model; next we need to fit it to the data, in order to obtain the estimates $\hat{\alpha}$,$\hat{\beta}$,$\hat{\gamma}$ and their standard errors.
- In `R`, logistic regression models are fitted using the `glm` function, as

```
model <- glm(cbind(dead,censored) ~ z + x,
             family=binomial(link="logit"))
```

- Here the outcome data were entered as frequency records.
- Alternatively, we could have entered the data as individual level records; the results would be equivalent (verify).

Results

```
model <- glm(cbind(dead,censored) ~ z + x,
             family=binomial(link="logit"))
print(summary(model), signif.stars = FALSE)

##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.952      1.004   -4.93  8.0e-07
## z             -0.182      0.159   -1.14     0.25
## x              4.778      1.011    4.73  2.3e-06
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 133.81237  on 3  degrees of freedom
## Residual deviance:   0.18471  on 1  degrees of freedom
## AIC: 20.74
##
## Number of Fisher Scoring iterations: 5
```

Confidence interval for the odds ratio

- As usual, we can transform a 95% confidence interval on the log-odds ratio scale to the odds ratio scale as

$$
\begin{aligned}
&e^{-0.1816 \pm 1.96 \times 0.1595} \\
&= 0.834 \times e^{\pm 0.313} \\
&= (0.610, 1.140)
\end{aligned}
$$

- The null value is included in the interval.
- After adjusting for age, these data do not give evidence against the null

$$\frac{\frac{\pi_{1X}}{1-\pi_{1X}}}{\frac{\pi_{0X}}{1-\pi_{0X}}} = 1$$

Logistic Regression

Log-linear model for risk

Kidney stone removal procedures 1

Kidney stone removal procedures 2

Diabetes cohort data 1

Diabetes cohort data 2

Log-linear model for risk

- Is there some particular reason why we *have* to use the logit link when modeling risk?
- Why could we not just parametrize the log-risk as

$$\log(\pi_{ZX}) = \alpha + \beta Z + \gamma X?$$

Log-linear model for risk

- Is there some particular reason why we *have* to use the logit link when modeling risk?
- Why could we not just parametrize the log-risk as

$$\log(\pi_{ZX}) = \alpha + \beta Z + \gamma X?$$

- We can; in this case the regression coefficient β would be interpreted as a log-risk ratio:

$$\begin{aligned}
\frac{\pi_{1X}}{\pi_{0X}} &= \frac{e^{\alpha+\beta+\gamma X}}{e^{\alpha+\gamma X}} \\
&= \frac{e^{\alpha} e^{\beta} e^{\gamma X}}{e^{\alpha} e^{\gamma X}} \\
&= e^{\beta} \\
\Leftrightarrow \log\left(\frac{\pi_{1X}}{\pi_{0X}}\right) &= \beta
\end{aligned}$$

Fitting the log-linear model

- To fit this model, we only need to change the link function:

```
model <- glm(cbind(dead,censored) ~ z + x,
             family=binomial(link="log"))
```

- Using the parameter estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$, risk estimates could then be obtained through the back-transformation

$$\hat{\pi}_{ZX} = e^{\hat{\alpha}+\hat{\beta}Z+\hat{\gamma}X}$$

- However, note that there is nothing here bounding the risk to values below one.
- The log-linear model does bound the risk to non-negative values, so as long as the risk is small, log-linear and logistic regression models give similar results.

Results

```
model <- glm(cbind(dead,censored) ~ z + x,
             family=binomial(link="log"))
print(summary(model), signif.stars = FALSE)

##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.967      0.997   -4.98  6.2e-07
## z             -0.102      0.089   -1.15     0.25
## x              4.182      0.999    4.19  2.8e-05
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 133.81237  on 3  degrees of freedom
## Residual deviance:   0.19892  on 1  degrees of freedom
## AIC: 20.76
##
## Number of Fisher Scoring iterations: 5
```

Interpretation

- The results from the log-linear model differ somewhat from the logistic model.
- This is unsurprising since the risk in the present example is not small, so we cannot approximate risk ratios by odds ratios.
- A 95% confidence interval for the risk ratio would be calculated in the usual way as

$$
\begin{aligned}
&e^{-0.10229 \pm 1.96 \times 0.08898} \\
&= 0.903 \times e^{\pm 0.174} \\
&= (0.759, 1.075)
\end{aligned}
$$

The 1986 BMJ article *Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy* by Charig et. al, was a study designed to compare different methods of treating kidney stones in order to establish which was the most cost effective and successful. The procedure, either open surgery, or percutaneous nephrolithotomy (PN, a keyhole surgery procedure), was defined to be successful if stones were eliminated or reduced to less than 2 mm after three months. The study collected cases of kidney stones treated at a particular UK hospital during 1972-1985. The counts of successes for the two surgical procedures were:

	Unsuccessful	Successful	Total
Open surgery	77	273	350
PN	61	289	350
Total	138	562	700

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.556      0.141  -11.04   <2e-16
## open           0.290      0.191    1.52     0.13
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.3148e+00  on 1  degrees of freedom
## Residual deviance: 9.9920e-15  on 0  degrees of freedom
## AIC: 15.7
##
## Number of Fisher Scoring iterations: 3
```

Below are the same outcomes tabulated by the size of the kidney stone (smaller than 2cm/at least 2cm in diameter):

< 2cm	Unsuccessful	Successful	Total
Open surgery	6	81	87
PN	36	234	270
Total	42	315	357

≥ 2cm	Unsuccessful	Successful	Total
Open surgery	71	192	263
PN	25	55	80
Total	96	247	343

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.937      0.170  -11.36  < 2e-16
## open          -0.357      0.229   -1.56     0.12
## size           1.261      0.239    5.27  1.3e-07
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.1239  on 3  degrees of freedom
## Residual deviance:  1.0082  on 1  degrees of freedom
## AIC: 26.36
##
## Number of Fisher Scoring iterations: 3
```

	Dead	Censored	Total
Type II	218	326	544
Type I	105	253	323
Total	323	579	902

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.879      0.116   -7.58  3.6e-14
## type           0.477      0.145    3.28    0.001
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0978e+01  on 1  degrees of freedom
## Residual deviance: 1.4033e-13  on 0  degrees of freedom
## AIC: 16.86
##
## Number of Fisher Scoring iterations: 2
```

Logistic Regression

Log-linear model for risk

Kidney stone removal procedures 1

Kidney stone removal procedures 2

Diabetes cohort data 1

Diabetes cohort data 2

Below are the same outcomes tabulated by age:

≤ 40	Dead	Censored	Total
Type II	0	15	15
Type I	1	129	130
Total	1	144	145

> 40	Dead	Censored	Total
Type II	218	311	529
Type I	104	124	228
Total	322	435	757

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.952      1.004   -4.93  8.0e-07
## type          -0.182      0.159   -1.14     0.25
## age            4.778      1.011    4.73  2.3e-06
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 133.81237  on 3  degrees of freedom
## Residual deviance:   0.18471  on 1  degrees of freedom
## AIC: 20.74
##
## Number of Fisher Scoring iterations: 5
```