BIOS 6312: Modern Biostatistics Methodology II

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 6: Binary outcome regression

Version: 04/26/2025

Table of Contents

Binary exposures:

- General setup of a 2×2 table (binary outcome and exposure):
  - X: 0 = unexposed; 1 = exposed.
  - Y: 0 = no disease; 1 = disease.

	$Y = 1$	$Y = 0$	Total
$X = 1$	a	b	$a + b$
$X = 0$	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Measures of outcome frequency:

- Risk* among exposed: $\mathrm{P}(Y = 1|X = 1)$, estimated as $(a/(a + b))$.
- Risk* among unexposed: $\mathrm{P}(Y = 1|X = 0)$, estimated as $(c/(c + d))$.
- Odds among exposed: $O(Y = 1|X = 1)$, estimated as a/b.
- Odds among unexposed: $O(Y = 1|X = 0)$ estimated as c/d.

*Note: Not every probability is a *risk*. Some are prevalences, proportions.

Binary exposures:

- General setup of a 2×2 table (binary outcome and exposure):
  - X: 0 = unexposed; 1 = exposed.
  - Y: 0 = no disease; 1 = disease.

	$Y = 1$	$Y = 0$	Total
$X = 1$	a	b	$a + b$
$X = 0$	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Back-of-the-envelope calculations of measures of association:

- Risk* difference (RD): $(a/(a + b)) - (c/(c + d))$.
- Odds ratio (OR): $(a \times d)/(b \times c)$.
- Risk* ratio (RR): $(a/(a + b))/(c/(c + d))$.

*Note: Not every probability is a *risk*. Some are prevalences, proportions.

Basic example: Diabetes and CHD

- As a motivating example, let us use the MRI cohort to examine the association between diabetes and coronary heart disease (CHD).
  - X: 0 = no diabetes; 1 = diabetes.
  - Y: 0 = no CHD; 1 = angina/myocardial infarction.

	CHD	No CHD	Total
Diabetes	23	56	79
No diabetes	132	524	656
Total	155	580	735

Example calculations:

- CHD prevalence among diabetes patients: $23/79 = 0.291 = 29.1\%$.
- Estimated odds ratio (OR): $(23 \times 524)/(56 \times 132) = 1.630$.

Table of Contents

Binary outcome regression

Setup: Binary outcomes

- So far in this course, we have considered both continuous and discrete exposures (and their respective interactions, etc.), but all of our outcomes have been treated continuously.
- Binary outcomes (e.g., death, myocardial infarction) are sometimes of interest as well.
  - $Y \sim \text{Bernoulli}(p)$; $0 < p < 1$.
  - $\text{P}(Y = y) = p^y(1-p)^{1-y}$.
    - ⋆ In other words, $\text{P}(Y = 0) = 1 - p$ and $\text{P}(Y = 1) = p$.
  - $\text{E}[Y] = 0 \times \text{P}(Y = 0) + 1 \times \text{P}(Y = 1) = \text{P}(Y = 1) = p$.
  - $\text{Var}[Y] = \overset{\text{math}}{\cdots} = p(1-p)$.
- Note relationship between mean and variance.
  - Variance maximal when $p = 0.5$; minimal when $p = 0$ or 1.

Binary outcome regression

Fist thought: Why not use prior approach?

- Goal: Evaluate extent to which mean of Y varies across X.
  - Note: When Y is binary, $\mathrm{E}[Y|X=x] = \mathrm{P}(Y=1|X=x)$.
- Linear model: $\mathrm{E}[Y|X=x] = \mathrm{P}(Y=1|X=x) = \beta_0 + \beta_1 x$.
  - $\beta_0 = \mathrm{P}(Y=1|X=0)$.
  - $\beta_1 = \mathrm{P}(Y=1|X=x+1) - \mathrm{P}(Y=1|X=x)$.
- Challenges:
  - A probability should be bounded by 0 and 1.
  - Linearity is therefore generally not plausible when X is continuous.
  - Known mean-variance relationship: $\mathrm{E}[Y] = p$; $\mathrm{Var}(Y) = p(1-p)$.
    - ⋆ Can be *addressed* by robust standard errors, but not leveraged.

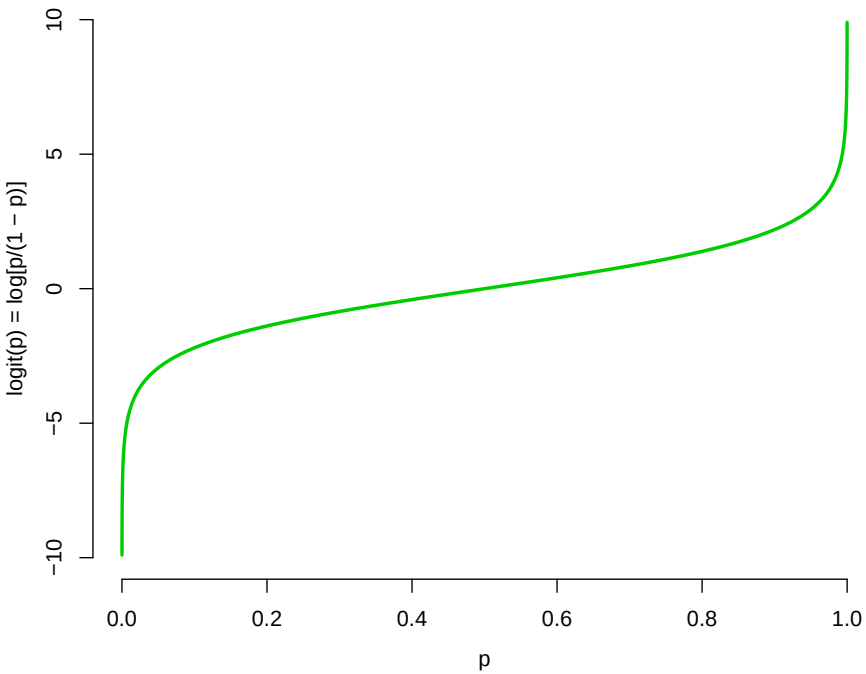
Other ideas:

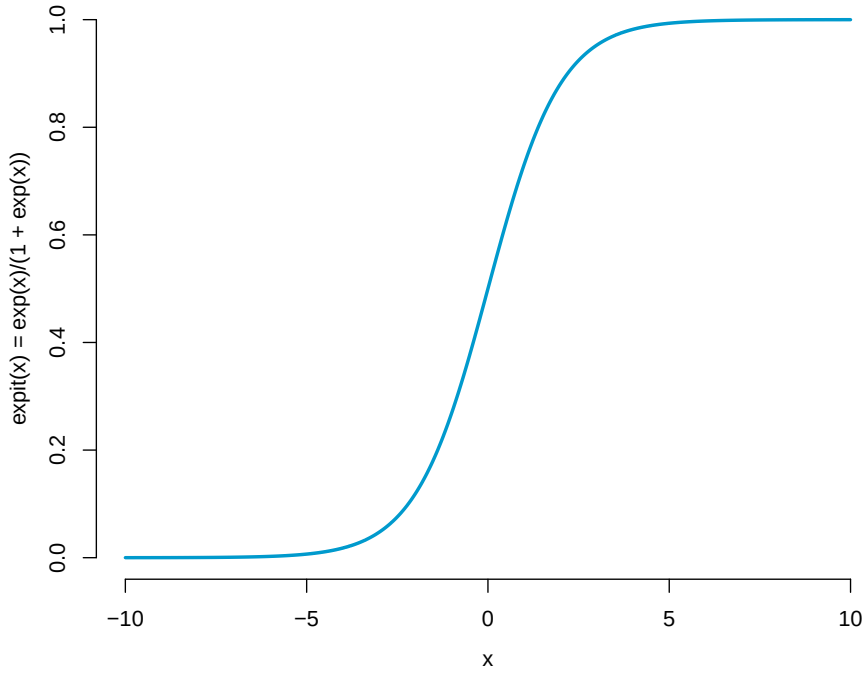
- Linear model: $\mathrm{E}[Y|X=x] = \mathrm{P}(Y=1|X=x) = \beta_0 + \beta_1 x$.
- Let g denote a function that will allow us to relate $\mathrm{P}(Y=1|X=x)$ to x. For a linear model, $g(p) = p$. In general, we would like:

$$\begin{aligned} g(\mathrm{P}(Y=1|X=x)) &= \beta_0 + \beta_1 x \\ \iff \mathrm{P}(Y=1|X=x) &= g^{-1}(\beta_0 + \beta_1 x). \end{aligned}$$

- One nice choice is $g(p) = \mathrm{logit}(p) = \log(p/(1-p))$.
  - $g : (0,1) \longrightarrow \mathbb{R}$, a nice property—we won't have to worry about probabilities escaping the range of possible values.
  - Note that $g^{-1}(x) = \mathrm{expit}(x) = e^x/(1+e^x)$.

Special functions: logit and expit





Simple logistic regression:

- Choosing $g(p) = \text{logit}(p)$: *logistic regression*.
- Model (the following are equivalent):

$$\begin{aligned}
\text{logit}(\text{P}(Y = 1|X = x)) &= \beta_0 + \beta_1 x. \\
\log\left(\frac{\text{P}(Y = 1|X = x)}{1 - \text{P}(Y = 1|X = x)}\right) &= \beta_0 + \beta_1 x. \\
\text{P}(Y = 1|X = x) &= \text{expit}(\beta_0 + \beta_1 x). \\
\frac{\text{P}(Y = 1|X = x)}{1 - \text{P}(Y = 1|X = x)} &= \exp(\beta_0 + \beta_1 x).
\end{aligned}$$

Simple logistic regression:

- Odds: $O(Y=1) = \mathrm{P}(Y=1)/[1-\mathrm{P}(Y=1)]$.
- Logistic regression model (the following are equivalent):

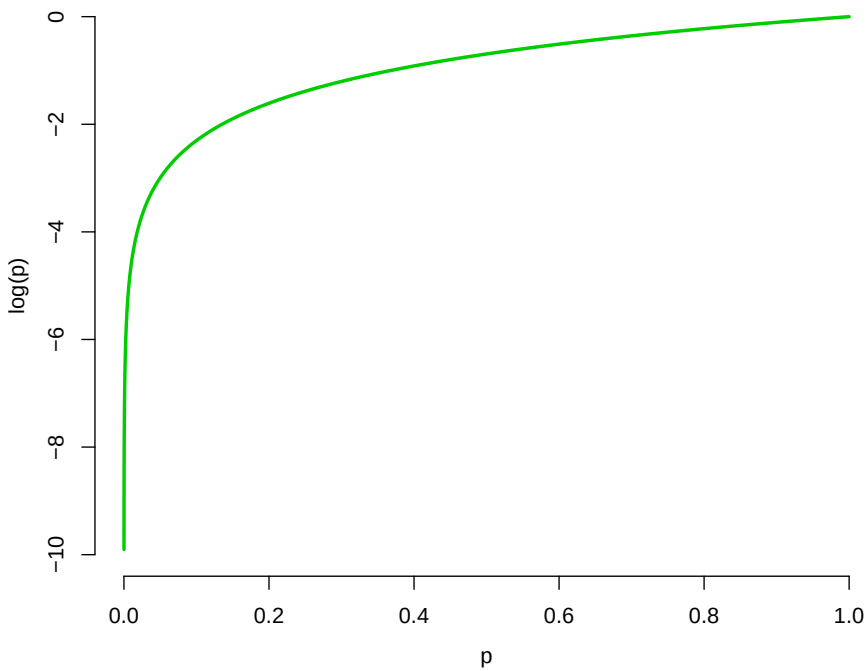
$$\begin{aligned} \log\left(O(Y=1|X=x)\right) &= \beta_0 + \beta_1 x. \\ O(Y=1|X=x) &= \exp(\beta_0 + \beta_1 x). \end{aligned}$$

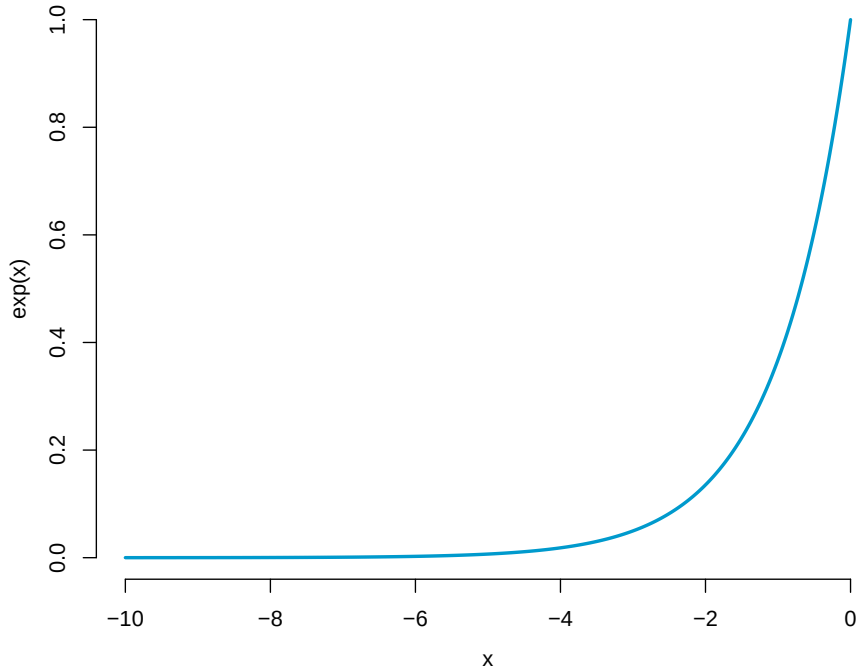
- Coefficient interpretation:
  - $\beta_0 = \log(O(Y=1|X=0))$.
  - $\beta_1 = \log(O(Y=1|X=x+1)) - \log(O(Y=1|X=x))$.
  - $\exp(\beta_0) = O(Y=1|X=0)$.
    - ★ Odds among stratum $X=0$.
  - $\exp(\beta_1) = O(Y=1|X=x+1)/O(Y=1|X=x)$.
    - ★ Odds ratio comparing strata differing in X by one unit.

Other ideas:

- So far: $g(p) = p$ (linear model) and $g(p) = \text{logit}(p)$ (logistic model).
- Yet another choice is $g(p) = \log(p)$.
  - $g : (0, 1) \longrightarrow (-\infty, 0)$; coefficients need constraints to avoid probabilities escaping range of possible values.
  - Note that $g^{-1}(x) = \exp(x)$.

Special functions: logarithm and exponential





Simple log-linear regression:

- Choosing $g(p) = \log(p)$: *relative risk* regression*.
- Model (the following are equivalent):

$$\begin{aligned} \log(\mathrm{P}(Y = 1|X = x)) &= \beta_0 + \beta_1 x. \\ \mathrm{P}(Y = 1|X = x) &= \exp(\beta_0 + \beta_1 x). \end{aligned}$$

- Coefficient interpretation:
  - $\beta_0 = \log(\mathrm{P}(Y = 1|X = 0))$.
  - $\beta_1 = \log(\mathrm{P}(Y = 1|X = x + 1)) - \log(\mathrm{P}(Y = 1|X = x))$.
  - $\exp(\beta_0) = \mathrm{P}(Y = 1|X = 0)$.
    - ⋆ Risk/prevalence among stratum $X = 0$.
  - $\exp(\beta_1) = \mathrm{P}(Y = 1|X = x + 1)/\mathrm{P}(Y = 1|X = x)$.
    - ⋆ Risk/prevalence ratio comparing strata differing in X by one unit.

*Keep in mind, not all probabilities are *risks*.

Summary so far:

- Linear regression of a binary outcome amounts to an analysis of difference in proportions. Particularly when the exposure is continuous, this turns out not to be a particularly good idea.
- Logistic regression (sometimes called log-odds regression), allows you to model the *log-odds* as a linear function of a predictor. From this, odds ratios can be recovered by exponentiating coefficients of interest.
- Log-linear regression (sometimes called relative risk regression), allows you to model the *log-risk/proportion/probability/prevalence*. From this, risk ratios can be recovered by exponentiating coefficients of interest.
- In this course, we'll tend to stick to the logistic model for binary outcomes.

Where we're headed:

- Understanding exactly how the parameters of a logistic/log-linear regression model are *estimated* is a great deal more challenging understanding estimation of a linear regression model.
- Many of the procedures and ideas from linear regression of continuous outcomes will generalize nicely to binary outcomes. Some, however, will not.
- Considerations regarding study design are critically important. Certain choices for $g(p)$ do not apply to certain study designs.
- For this class, don't worry too much about how the numbers are being estimated (the topic of an advanced regression course).

Estimation:

- Recall that estimators for linear regression models possessed closed-form expressions (none of which I need you to memorize).
- Linear, logistic, and log-linear models are all examples of *generalized linear models*. There is a procedure for figuring out the equations to estimate $\boldsymbol{\beta}$; the theory of their derivation is outside the scope of this class. In general, the equations take the following form:

$$\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0},$$

where:

  - $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ represents $\mathrm{E}[\mathbf{y}|\mathbf{X}]$.
  - $\mathbf{D} = \mathbf{D}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$ is a derivative matrix.
  - $\mathbf{V} = \mathbf{V}(\boldsymbol{\beta})$ is a (diagonal) variance matrix.

Estimation:

- Estimating the parameters of a generalized linear model (GLM) invariably involves solving the following equations.

$$\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{0},$$

- It doesn't look too far off from the equations that were used to solve the OLS equations, once you get away from a linear choice for $g(\cdot)$, the equations generally don't possess a closed-form solution.
- However, for a given setting (e.g., binary outcomes), there is generally a mathematically "nice" choice for $g(\cdot)$ that makes the math work nicely. This choice is called the canonical link function.
  - For binary outcomes, the canonical link function is the logit function.
- Newton-Raphson methods can be used to find numerical solutions. Software will do that for you!

Binary outcome regression

Variance:

- Recall ideas of model-based and model-agnostic variance estimation for regression of continuous outcomes.
  - By "model-based", I specifically mean model based in reference to the mean-variance relationship in this case.
  - For OLS, the model-based variance presumes homoscedasticity and the robust/sandwich variance (model-agnostic) allows heteroscedasticity.
- The sandwich variance extends to logistic/log-linear regression of binary outcomes, though it isn't model-agnostic in the same way.
- For (independent) binary outcomes, it is impossible to decouple the mean and variance. Therefore, specifying a link function implicitly specifies a mean-variance relationship.

Testing:

- Model: $g(\mathrm{E}[Y|X = x)) = \beta_0 + \beta_1 x$.
- Asymptotic behavior:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathrm{SE}}(\widehat{\beta}_1)} \dot{\sim} \mathcal{N}(0, 1).$$

- For logistic regression, β_1 corresponds to a difference in log-odds, and in relative risk regression, β_1 corresponds to a difference in log-risk.
- To obtain confidence intervals for odds/risk ratios, we typically exponentiate the endpoints of the confidence intervals for β_1.
  - The transformed confidence intervals are not symmetric.

Example: CHD and diabetes in MRI study cohort

- Let us use the MRI study to examine the association between diabetes and CHD.
  - X: 0 = no diabetes; 1 = diabetes.
  - Y: 0 = no CHD; 1 = angina/myocardial infarction.

	CHD	No CHD	Total
Diabetes	23	56	79
No diabetes	132	524	656
Total	155	580	735

- Estimated odds ratio (OR): 1.630.
- Estimated prevalence ratio (RR): 1.447.

Example: CHD and diabetes in MRI cohort

```
## Read in data
dat <- read.csv("mri.csv")

## Re-code CHD as a binary variable for this set of notes
dat$chd2 <- as.numeric(dat$chd > 0)

## Fit model (abridged output)
model <- regress("odds", chd2 ~ diabetes, data = dat)

> model

Call:
  regress(fnctl = "odds", formula = chd2 ~ diabetes, data = dat)

Deviance Residuals:
  Min      1Q  Median      3Q     Max
-0.83   -0.67   -0.67   -0.67    1.79

Coefficients:

  Raw Model:
                Estimate   Naive SE  Robust SE   F stat  df     Pr(>F)
[1] Intercept      -1.38     0.0974     0.0975   199.87   1  < 0.00005
[2] diabetes       0.489      0.266      0.266     3.36   1      0.067

  Transformed Model:
                  e(Est)    e(95%L)   e(95%H)   F stat  df.    Pr(>F)
[1] Intercept      0.252      0.208     0.305   199.87   1  < 0.00005
[2] diabetes        1.63      0.966      2.75     3.36   1      0.067
```

Example: CHD and diabetes in MRI cohort

- Whether you prefer the untransformed or transformed output will depend upon the scenario.
- Indeed, both the logistic and log-linear models are *saturated* in this example (in which the exposure is also binary). They encode within them no assumptions of linearity on any scale.

Other considerations: Model building

- Many ideas behind model building carry directly over to the setting of binary outcome regression.
- Let us, as an example, build a logistic model for CHD based on diabetes (0 = no, 1 = yes), continuous age (years), and an interaction between between the two.
- Variables:
  - Y: CHD (0 = none, 1 = angina/myocardial infarction)
  - X_1: diabetes (0 = no, 1 = yes).
  - X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- As a quick exercise, let's interpret the model coefficients.
- Try $\exp(\beta_0)$, $\exp(\beta_1)$, and $\exp(\beta_2)$!

Example: CHD and diabetes in MRI cohort

- Variables:
  - Y: CHD (0 = none, 1 = angina/myocardial infarction)
  - X_1: diabetes (0 = no, 1 = yes).
  - X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- β_3 is the most challenging!

$$\begin{aligned}\beta_3 &= [\log(O(Y = 1|X_1 = x_1 + 1, X_2 = x_2 + 1)) - \log(O(Y = 1|X_1 = x_1, X_2 = x_2 + 1))] \\ &\quad - [\log(O(Y = 1|X_1 = x_1 + 1, X_2 = x_2)) - \log(O(Y = 1|X_1 = x_1, X_2 = x_2))]\end{aligned}$$

$$\exp(\beta_3) = \frac{O(Y = 1|X_1 = x_1 + 1, X_2 = x_2 + 1)/O(Y = 1|X_1 = x_1, X_2 = x_2 + 1)}{O(Y = 1|X_1 = x_1 + 1, X_2 = x_2)/O(Y = 1|X_1 = x_1, X_2 = x_2)}$$

- $\exp(\beta_3)$ represents a ratio of odds ratios, much the way an interaction term in a linear model represents a difference in mean differences.

Example: CHD and diabetes in MRI cohort (`logit`)

```
## Fit model (abridged output)
model <- regress("odds", chd2 ~ diabetes*age, data = dat)

> model

Call:
  regress(fnctl = "odds", formula = chd2 ~ diabetes * age, data = dat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.920  -0.700  -0.644  -0.610   1.897

Coefficients:

  Raw Model:
                    Estimate  Naive SE  Robust SE  F stat  df    Pr(>F)
[1] Intercept         -3.67      1.29       1.21    9.18.  1    0.0025
[2] diabetes           3.24      3.73       3.61    0.80   1    0.3703
[3] age              0.0306    0.0172     0.0161    3.61.  1    0.0577
[4] diabetes:age    -0.0368    0.0500     0.0485    0.58   1    0.4485

  Transformed Model:
                      e(Est)    e(95%L)    e(95%H)   F stat  df    Pr(>F)
[1] Intercept         0.0255    0.00237     0.275      9.18   1    0.0025
[2] diabetes            25.5     0.0212     30603      0.80   1    0.3703
[3] age                 1.03      0.999      1.06      3.61   1    0.0577
[4] diabetes:age       0.964      0.876      1.06      0.58   1    0.4485
```

Example: CHD and diabetes in MRI cohort

- Using the tools we've already developed throughout the course, we can run through a number of examples from this one model. This will serve as a nice review!
  1. Characterize the strength of evidence of an overall age-adjusted association between diabetes and CHD.
  2. Construct a point estimate and 95% confidence interval for the odds ratio that compares the odds of CHD between non-diabetic adults differing in their age by five years.
  3. Construct a point estimate and 95% confidence interval for the odds of CHD among non-diabetics of age 70.
  4. Construct a point estimate and 95% confidence interval for the prevalence of CHD among diabetics of age 82.
  5. Construct a point estimate and 95% confidence interval for the odds ratio that compares the odds of CHD between diabetic and non-diabetic adults of age 73.

Example: CHD and diabetes in MRI cohort

- Y: CHD (0 = none, 1 = angina/myocardial infarction)
- X_1: diabetes (0 = no, 1 = yes).
- X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- **Example 1**: Characterize the strength of evidence of an overall age-adjusted association between diabetes and CHD.
  - The null hypothesis is represented as $H_0 : \beta_1 = \beta_3 = 0$.

Example: CHD and diabetes in MRI cohort (`logit`)

```
## Reminder of model
model <- regress("odds", chd2 ~ diabetes*age, data = dat)

## Constraint matrix
R <- matrix(0, nrow = 2, ncol = 4)
R[1,2] <- R[2,4] <- 1

## Test of interest
> lincom(model, R, joint.test = TRUE)
     Chi2 stat df p value
[1,]     4.168  2   0.124
```

Example: CHD and diabetes in MRI cohort

- Y: CHD (0 = none, 1 = angina/myocardial infarction)
- X_1: diabetes (0 = no, 1 = yes).
- X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- **Example 2**: Construct a point estimate and 95% confidence interval for the odds ratio that compares the odds of CHD between non-diabetic adults differing in their age by five years.
  - The sub-model for non-diabetics is given by:

$$\log(O(Y = 1|X_1 = 0, X_2 = x_2)) = \beta_0 + \beta_2 x_2.$$

  - Adults differing in their age by five years should therefore differ in their log-odds by $5\beta_2$. Therefore, the odds ratio of interest is given by $\exp(5\beta_2) = [\exp(\beta_2)]^5$.

Example: CHD and diabetes in MRI cohort)

```
## Reminder of model
model <- regress("odds", chd2 ~ diabetes*age, data = dat)

## Constraint matrix
R <- matrix(c(0,0,5,0), nrow = 1, ncol = 4)

## Test of interest
> lincom(model, R)

H0: 5*age   =  0
Ha: 5*age  !=  0
      e(Est) Std. Err. e(95%L) e(95%H)     T Pr(T > |t|)
[1,] 1.16516   0.08042 0.99499 1.36444 1.901      0.0577
```

Example: CHD and diabetes in MRI cohort

- Y: CHD (0 = none, 1 = angina/myocardial infarction)
- X_1: diabetes (0 = no, 1 = yes).
- X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- **Example 3**: Construct a point estimate and 95% confidence interval for the odds of CHD among non-diabetics of age 70.
  - The log-odds in question is given by:

$$\log(O(Y = 1|X_1 = 0, X_2 = 70)) = \beta_0 + 70\beta_2.$$

  - We can back-transform out of this:

$$O(Y = 1|X_1 = 0, X_2 = 70) = \exp(\beta_0 + 70\beta_2).$$

Example: CHD and diabetes in MRI cohort (`logit`)

```
## Reminder of model
model <- regress("odds", chd2 ~ diabetes*age, data = dat)

## Constraint matrix
R <- matrix(c(1,0,70,0), nrow = 1, ncol = 4)

## Test of interest
> lincom(model, R)

H0: 1*(Intercept)+70*age   =  0
Ha: 1*(Intercept)+70*age  !=  0
     e(Est) Std. Err. e(95%L) e(95%H)       T Pr(T > |t|)
[1,] 0.2170    0.1269  0.1692  0.2784 -12.04      <2e-16 ***
```

Example: CHD and diabetes in MRI cohort

- Y: CHD (0 = none, 1 = angina/myocardial infarction)
- X_1: diabetes (0 = no, 1 = yes).
- X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- **Example 4**: Construct a point estimate and 95% confidence interval for the prevalence of CHD among diabetics of age 82.
  - The log-odds in question is given by:

$$\log(O(Y = 1|X_1 = 1, X_2 = 82) = \beta_0 + \beta_1 + 82(\beta_2 + \beta_3).$$

  - We can back-transform out of this:

$$\mathrm{P}(Y = 1|X_1 = 0, X_2 = 82) = \mathrm{expit}(\beta_0 + \beta_1 + 82(\beta_2 + \beta_3)).$$

- You need to back out of the transformation yourself.

Example: CHD and diabetes in MRI cohort (`logit`)

```
## Reminder of model
model <- regress("odds", chd2 ~ diabetes*age, data = dat)

## Constraint matrix
R <- matrix(c(1,1,82,82), nrow = 1, ncol = 4)

## Test of interest
> lincom(model, R, eform = FALSE)

H0: 1*(Intercept)+1*diabetes+82*age+82*diabetes:age   =  0
Ha: 1*(Intercept)+1*diabetes+82*age+82*diabetes:age  !=  0
     Estimate Std. Err.     95%L     95%H      T Pr(T > |t|)
[1,] -0.93812   0.43510 -1.79231 -0.08392 -2.156      0.0314 *

## Note: Apply expit() to (-1.79231) and (-0.08392)
## and obtain [0.143, 0.478] as a 95% CI for
## the probability of interest
```

Example: CHD and diabetes in MRI cohort

- Y: CHD (0 = none, 1 = angina/myocardial infarction)
- X_1: diabetes (0 = no, 1 = yes).
- X_2: continuous age (years).
- Model: $\log(O(Y = 1|X_1 = x_1, X_2 = x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- **Example 5**: Construct a point estimate and 95% confidence interval for the odds ratio that compares the odds of CHD between diabetic and non-diabetic adults of age 73.
  - The sub-model for those of age 73 years is given by:

$$\log(O(Y = 1|X_1 = x_1, X_2 = 73)) = \beta_0 + \beta_1 x_1 + 73(\beta_2 + \beta_3 x_1).$$

  - The difference in log-odds is therefore given by $\beta_1 + 73\beta_3$; the odds ratio of interest can be obtained by exponentiating.

Example: CHD and diabetes in MRI cohort (`logit`)

```
## Reminder of model
model <- regress("odds", chd2 ~ diabetes*age, data = dat)

## Constraint matrix
R <- matrix(c(0,1,0,73), nrow = 1, ncol = 4)

## Test of interest
> lincom(model, R)

H0: 1*diabetes+73*diabetes:age   =  0
Ha: 1*diabetes+73*diabetes:age  !=  0
     e(Est) Std. Err. e(95%L) e(95%H)     T Pr(T > |t|)
[1,] 1.7397    0.2744  1.0151  2.9814 2.018        0.044 *
```

Diagnostics:

- Diagnostic plots and tests are possible in binary outcome regression.
- One simple possibility is to utilize standardized residuals:

$$r_i = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}},$$

  where, for logistic regression, $\widehat{\pi}_i = \text{expit}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}})$.
- Standardized residuals do not take into account variability due to estimation of $\boldsymbol{\beta}$, and will also generally not be approximately normally distributed. However, they should have an *approximate* mean of zero and variance of one.
- Standardized residuals with $|r_i| > 2$ may indicate lack of model fit.

Diagnostics:

- Other methods are available:
  - Deviance residuals.
  - Pregibon leverage.
  - Hosmer-Lemeshow goodness-of-fit.
- It is important to propose models that are sensible. In general, evidence of modest departures should not immediately cause you to distrust the conclusions of your model.
- In addition, always remember that diagnostics cannot prove to you that a model is correct.

Binary outcome regression

Example: CHD and diabetes in MRI cohort

- Hopefully, this example underscores that many of the "regression math" procedures generalize quite nicely.
- The reason for this is that generalized linear models are still linear on some scale (log-odds, log-risk, etc.). This is what allows us to use the same manipulations.
- In order to obtain clinically meaningful estimates, you need to back-transform.
  - For logistic regression, exponentiating coefficients or combinations of coefficients will generally produce stratum-specific odds or odds ratios. Back-transforming with the expit function puts you on the probability scale.

Additional examples:

- Procedures readily generalize:
  - Interaction terms and stratum-specific associations.
  - Categorical predictors.
  - Shifting/scaling, and log-transformation of of predictors.
  - Basis expansions/splines.
- Be able to navigate these ideas for non-continuous outcomes, even if we don't go over every possible case in every possible setting. Let's take a couple more examples together.
- **Example 1**: Suppose X is a positive-valued predictor.
  - Model: $\text{logit}(\text{P}(Y = 1|X = x)) = \beta_0 + \beta_1 \log(x)$.
  - Exercise: Interpret $(1 + q)^{\beta_1}$.
- **Example 2**: Suppose X is a categorical predictor $(X = 0, 1, 2)$.
  - Model: $\log(\text{P}(Y = 1|X = x)) = \beta_0 + \beta_1 1(x = 1) + \beta_2 1(x = 2)$.
  - Exercise: Determine an expression for the risk ratio that compares the risk of $Y = 1$ between the strata $X = 2$ and $X = 1$.

Table of Contents

Outcome-dependent sampling

Considerations regarding study design:

- Note that not all measures of frequency and/or association can be estimated in all study designs.
- Case-control studies employ outcome-dependent sampling. As such, they cannot estimate the outcome prevalence.
  - Interestingly, odds ratios are estimable in case-control studies via some mathematical hocus-pocus.
- The table below summarizes what population parameters can be estimated under various study designs without external information.

	$P(X=1)$	$P(Y=1)$	$P(Y=1\|X=1)$	RD	RR	OR
Randomized-controlled trial	n/a	N	Y	Y	Y	Y
Exposure-dependent cohort	N	N	Y	Y	Y	Y
Simple random sample	Y	Y	Y	Y	Y	Y
Case-control study	N	N	N	N	N	Y

Outcome-dependent sampling

Sampling schemes: An important note

- Cohort study: Sample by exposure.
  - Sample $N_0 = 500$ smokers and $N_1 = 500$ nonsmokers.
  - May estimate risk of cancer in each stratum.
- Case-control study: Sample by outcomes.
  - Sample $N_0 = 500$ controls and $N_1 = 500$ cancer patients.
  - May estimate prevalence of smoking in each stratum.
  - Most typical when outcome is rare.

Outcome-dependent sampling

Simple logistic regression: Case-control studies

- Case-control studies use *outcome-dependent* sampling.
- Under this study design, $\mathrm{P}(Y = 1)$ cannot be estimated.
  - In turn, conditional probability $\mathrm{P}(Y = 1|X = x)$ cannot be estimated, nor can the corresponding unconditional and conditional odds—$O(Y = 1)$ and $O(Y = 1|X = x)$.
- Mathematical magic (easiest to prove when X is binary):

$$\frac{O(Y = 1|X = x + 1)}{O(Y = 1|X = x)} = \frac{O(X = x|Y = 1)}{O(X = x|Y = 0)},$$

- RHS *can* be identified $\Rightarrow$ LHS can be too.
  - Odds ratios can be estimated in case-control studies.
- In an ideal world, we get to choose target parameter—but the nature of the data can sometimes tie our hands: no relative risk regression in case-control studies.

Simple logistic regression: Case-control studies

- Model: $\log\left(O(Y=1|X=x)\right) = \beta_0 + \beta_1 x$.
- Coefficient interpretation:
  - $\exp(\beta_0) = O(Y=1|X=0)$: Odds among stratum $X=0$.
    - ★ Cannot be identified in case-control study due to biased sampling of the outcome.
  - $\exp(\beta_1) = O(Y=1|X=x+1)/O(Y=1|X=x)$: Odds ratio comparing strata differing in X by one unit.
    - ★ Can be identified in case-control study!

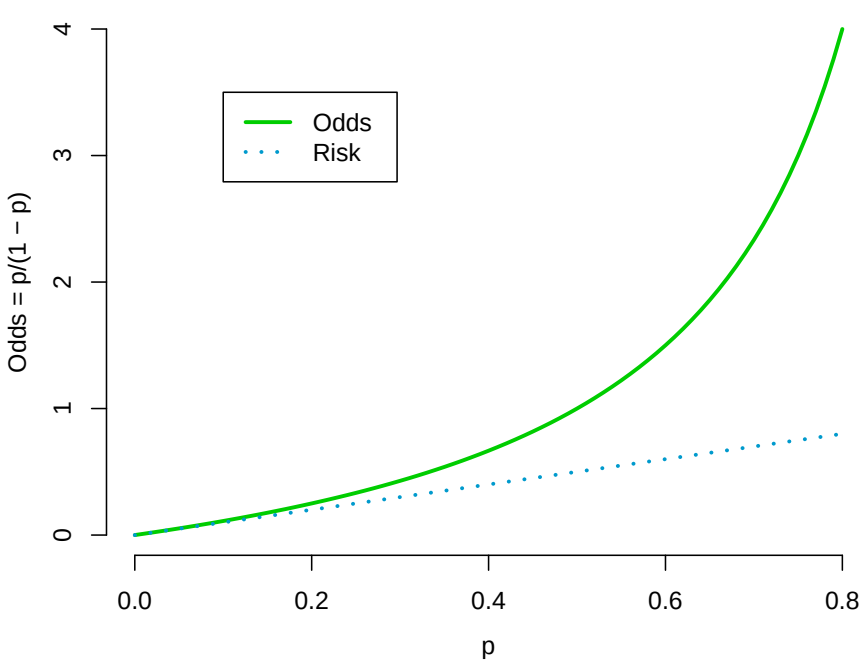
Rare outcomes:

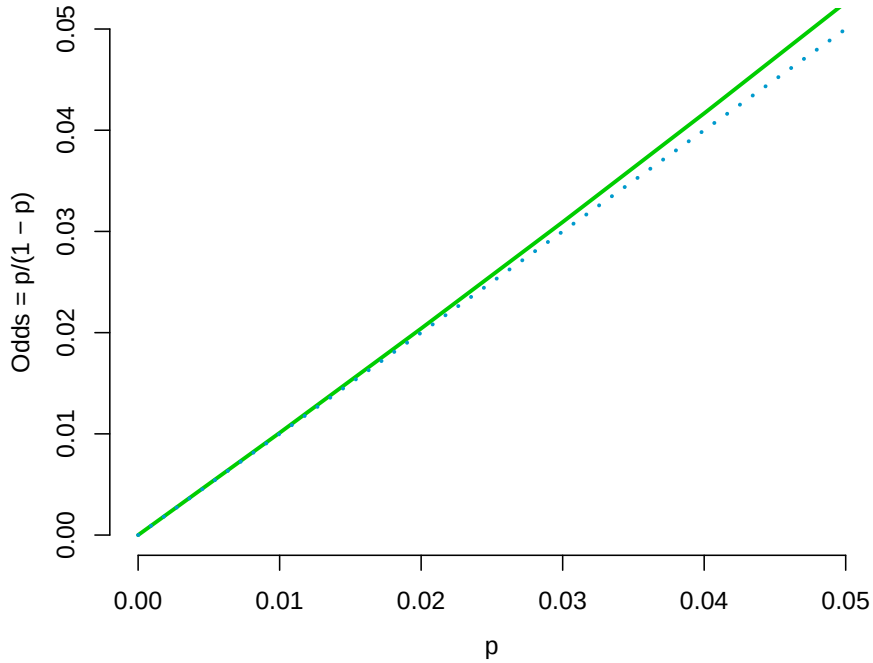
- If Y has low prevalence in the population (i.e., if $\mathrm{P}(Y = 1)$ is low), then the odds approximates the risk/prevalence:

$$\frac{\mathrm{P}(Y = 1)}{1 - \mathrm{P}(Y = 1)} \approx \mathrm{P}(Y = 1).$$

- In such settings, the odds ratio can be interpreted as an approximation to the risk ratio.
  - This is true regardless of the sampling scheme, though the most common reason to use a case-control study is a low outcome prevalence.

Rare outcomes: Odds as an approximation to risk





Outcome-dependent sampling

Example: Esophageal cancer study

- A case-control study was conducted to evaluate risk factors for esophageal cancer (data: `esoph.csv`).
  - Controls (no esophageal cancer): $N_0 = 975$
  - Cases (esophageal cancer): $N_1 = 200$
- One of the risk factors in question was alcohol consumption:
  - 1: 0-39 g/day
  - 2: 40-79 g/day
  - 3: 80-119 g/day
  - 4: 120+ g/day
- Because these data arise from a case-control study, we are quite limited in the ways we can characterize an association between the predictor and the outcome. Logistic regression can be used to estimate odds ratios.

Outcome-dependent sampling

Example: Esophageal cancer study

```
> table(dat$alcgrp, dat$esophcancer)

     0   1
1 415  29
2 355  75
3 138  51
4  67  45
```

Outcome-dependent sampling

Example: Esophageal cancer study

- Variables:
  - Y: esophageal cancer (0 = no, 1 = yes).
  - X: alcohol, g/day (1 = 0-39, 2: 40-79, 3: 80-119, 4: 120+).
- Model: $\log(O(Y = 1|X = x)) = \beta_0 + \beta_1 1(x = 2) + \beta_2 1(x = 3) + \beta_3 1(x = 4)$.
- Understanding the coefficients:
  - $\exp(\beta_0)$: odds of esophageal cancer among those with an alcohol consumption of 0-39 g/day.
    - ⋆ An estimate of this is meaningless in this study because participants with esophageal cancer are (highly) over-represented in the study.
  - $\exp(\beta_1)$: odds ratio, comparing the odds of esophageal cancer between those with an alcohol consumption of 40-79 g/day and those with an alcohol consumption of 0-39 g/day.
    - ⋆ This is estimable!
  - Etc., etc.

Example: Esophageal cancer study

```
## This is a categorical variable
dat$alcgrp <- factor(dat$alcgrp)

## Fit model (abridged output - only showing transformed)
model <- regress("odds", esophcancer ~ alcgrp, data = dat)

> model

Call:
  regress(fnctl = "odds", formula = esophcancer ~ alcgrp, data = dat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.014  -0.619  -0.367  -0.367   2.336

Coefficients:

  Transformed Model:
                   e(Est)   e(95%L)   e(95%H)   F stat  df     Pr(>F)
[1] Intercept      0.0699    0.0479    0.102    191.28   1  < 0.00005
alcgrp                                          25.81    3  < 0.00005
[2]    2            3.02      1.92      4.75     23.00.  1  < 0.00005
[3]    3            5.29      3.22      8.69     43.37   1  < 0.00005
[4]    4            9.61      5.63      16.4     68.93.  1  < 0.00005
```

Outcome-dependent sampling

Example: Esophageal cancer study

- Variables: alcohol consumption:
  - Y: esophageal cancer ($0 =$ no, $1 =$ yes).
  - X: alcohol, g/day ($1 =$ 0-39, 2: 40-79, 3: 80-119, 4: 120+).
- Model: $\log(O(Y = 1|X = x)) = \beta_0 + \beta_1 1(x = 2) + \beta_2 1(x = 3) + \beta_3 1(x = 4)$.
- Suppose we seek an expression for the odds ratio that compares the odds of esophageal cancer between those with an alcohol consumption of 120+ g/day and those with an alcohol consumption of 40-79 g/day.
  - With a little regression math, you should find that this is given by $\exp(\beta_3 - \beta_1)$.

Example: Esophageal cancer study

```
## Reminder of model
model <- regress("odds", esophcancer ~ alcgrp, data = dat)

## Constraint matrix
R <- matrix(c(0,-1,0,1), nrow = 1, ncol = 4)

## Test of interest
> lincom(model, R)

H0: -1*alcgrp2+1*alcgrp4   =  0
Ha: -1*alcgrp2+1*alcgrp4  !=  0
     e(Est) Std. Err. e(95%L) e(95%H)     T Pr(T > |t|)
[1,] 3.1791    0.2313  2.0196  5.0044 5.001    6.56e-07 ***

## Because esophageal cancer is rare in the population,
## I can present this as an approximate estimate of the
## corresponding risk ratio.
```

Outcome-dependent sampling

Example: Esophageal cancer study

- Variables:
  - Y: esophageal cancer ($0 =$ no, $1 =$ yes).
  - X: alcohol, g/day ($1 =$ 0-39, 2: 40-79, 3: 80-119, 4: 120+).
- Model: $\log(O(Y = 1|X = x)) = \beta_0 + \beta_1 1(x = 2) + \beta_2 1(x = 3) + \beta_3 1(x = 4)$.
- We can evaluate whether there is an overall association between alcohol consumption and esophageal cancer by conducting an omnibus test.
  - Null hypothesis is given by $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Example: Esophageal cancer study

```
## Reminder of model
model <- regress("odds", esophcancer ~ alcgrp, data = dat)

## Constraint matrix
R <- matrix(0, nrow = 3, ncol = 4)
R[1,2] <- R[2,3] <- R[3,4] <- 1

## Test of interest
> lincom(model, R, joint.test = TRUE)
     Chi2 stat df p value
[1,]     77.42  3  <2e-16 ***

## Wald based chi-square test, which is asymptotically
## equivalent to Pearson's chi-square (score test)
```

Additional thoughts:

- We have seen that a lot of the procedures generalize quite nicely to binary outcome regression.
- Being able to manipulate/interpret parameters is important.
- It is equally important to know what can be estimated and when.
- Checking for understanding:
  1. Can you generally use logistic regression in a cohort study?
  2. Can you generally use log-linear regression in a case-control study?
  3. Can a logistic regression model be used to easily characterize stratum-specific risk in a cohort study?
  4. Can a logistic regression model be used to easily characterize risk ratios in a cohort study?

Table of Contents

Confounding

Basic ideas:

- When considering binary outcomes, confounding can pose challenges that are analogous to those we saw when considering continuous outcomes.
- Regardless of study design (cross-sectional, cohort, case-control), and regardless of whether you're modeling odds ratio or risk ratio, confounding *must* be considered and addressed.
  - Systematic confounding does not pose barriers in randomized trials.

Example: Smoking and lung cancer

- Consider a case-control study to examine the association between smoking and lung cancer.

	Lung cancer	No lung cancer	Total
Smoker	55	30	85
Not smoker	50	150	200
Total	105	180	285

- Estimated odds ratio: $\widehat{\text{OR}} = 5.50$.
  - Sometimes called *crude* odds ratio (unadjusted).

Example: Smoking and lung cancer

- Broken down by age group:

Age > 60	Lung cancer	No lung cancer
Smoker	50	25
Not smoker	25	50

Age ≤ 60	Lung cancer	No lung cancer
Smoker	5	5
Not smoker	25	100

- Age is clearly associated with both smoking and lung cancer.

Example: Smoking and lung cancer

- Broken down by age group:

Age > 60	Lung cancer	No lung cancer
Smoker	50	25
Not smoker	25	50

Age ≤ 60	Lung cancer	No lung cancer
Smoker	5	5
Not smoker	25	100

- Estimated odds ratio within each group: $\widehat{\text{OR}} = 4.00$.
  - *Adjusted* odds ratio.

Table of Contents

Collapsibility

Basic ideas:

- Consider a variable, Z, associated with Y but not with X.
  - In linear regression, referred to Z as a *precision* variable as it reduced variance without changing the value of the parameter being estimated.
  - This happened because mean differences are *collapsible*.
- This doesn't always happen. Some measures are what we refer to as *non-collapsible*, implying that adjustment for a variable like Z will change the value of the parameter being estimated.
  - Risk ratios are *collapsible*.
  - Odds ratios are *not collapsible*.
- Let's see how this works with a couple of straightforward examples.

Example: Phase I RCT

- Stratified by age group:

Older	Death/progression	No death/progression
Chemotherapy	4	6
No chemotherapy	8	2

Younger	Death/progression	No death/progression
Chemotherapy	3	7
No chemotherapy	6	4

- Age associated with risk of death/progression, but not with probability of chemotherapy.
  - Is age a confounder? No!!
- Adjusted (within-group) risk ratios: $\widehat{\text{RR}} = 0.500$.

Example: Phase I RCT

	Death/progression	No death/progression	Total
Chemotherapy	7	13	20
No chemotherapy	14	6	20
Total	21	19	40

- If I collapse tables over the age categories, the crude risk ratio does not differ from the stratum-specific risk ratios.
- Crude risk ratio: $\widehat{\text{RR}} = 0.500$.

Example: Cohort study of CVD and kidney stones

- Stratify by age group:

≤ 50	CVD	No CVD
Kidney stones	40	60
No kidney stones	10	90

> 50	CVD	No CVD
Kidney stones	90	10
No kidney stones	60	40

- Age group associated with CVD, but not kidney stones.
  - Is age a confounder? No!!
- Adjusted (within-group) odds ratios: $\widehat{\text{OR}} = 6.0$.

Collapsibility

Example: Cohort study of CVD and kidney stones

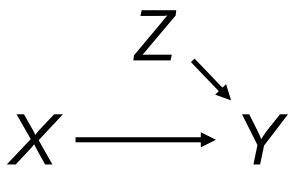
	CVD	No CVD	Total
Kidney stones	130	70	200
No kidney stones	70	130	200
Total	200	200	400

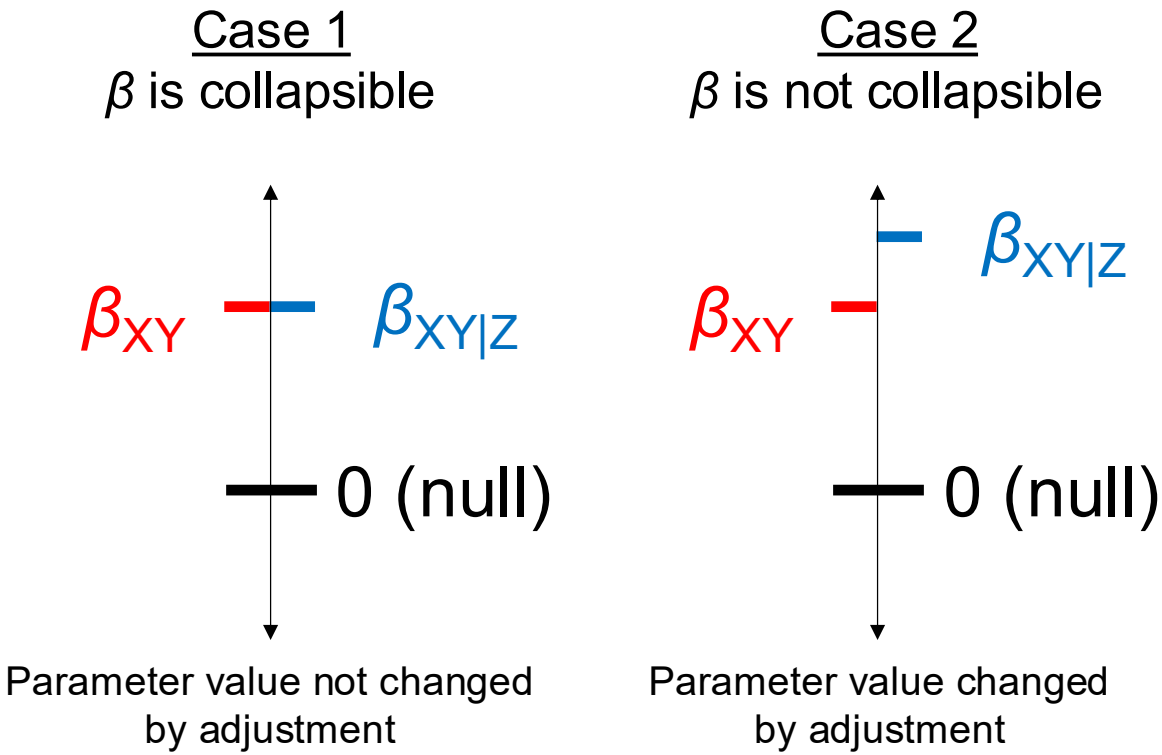
- If I collapse tables over the age categories, the crude odds ratio differs from the stratum-specific odds ratios.
- Crude odds ratio: $\widehat{\text{OR}} = 3.45$.

Adjustment in binary outcome regression:

- Now, reverse the logic: if we perform regression with a non-collapsible measure, then *adjustment* for a variable, Z, will fundamentally change the value of the quantity we're estimating. More often than not, the adjusted parameter will be further away from the null.
- A factor, Z, is best adjusted for when it is associated with Y (outcome), even if it is unrelated to X (exposure).
  - When the measure of association is collapsible, the justification is to gain precision.
  - When the measure of association is not collapsible, the justification is that the *crude* association is typically closer to the null than the stratum-specific association.
  - Not making a general statement about precision gains for non-collapsible links.

Collapsibility





Summary

This unit:

- Why not to use linear regression for binary outcomes.
- Logistic (primarily) and log-linear models for binary outcomes.
- Odds ratios (primarily), risk ratios, and regression math.
- Considerations for outcome-dependent sampling.
- Confounding.
- Collapsibility (risk ratios are collapsible; odds ratios are not).

Summary

So far:

- Review.
- Simple linear regression.
- Multiple linear regression (foundations).
- Multiple linear regression (interactions and strata).
- Transformations and basis expansions.
- Regression with binary outcomes.

Summary

Coming up:

- Regression with nominal, ordinal, and count outcomes.
- Introduction to clustered data.
- Methods for time-to-event outcomes.
- Predictive capacity of regression models.