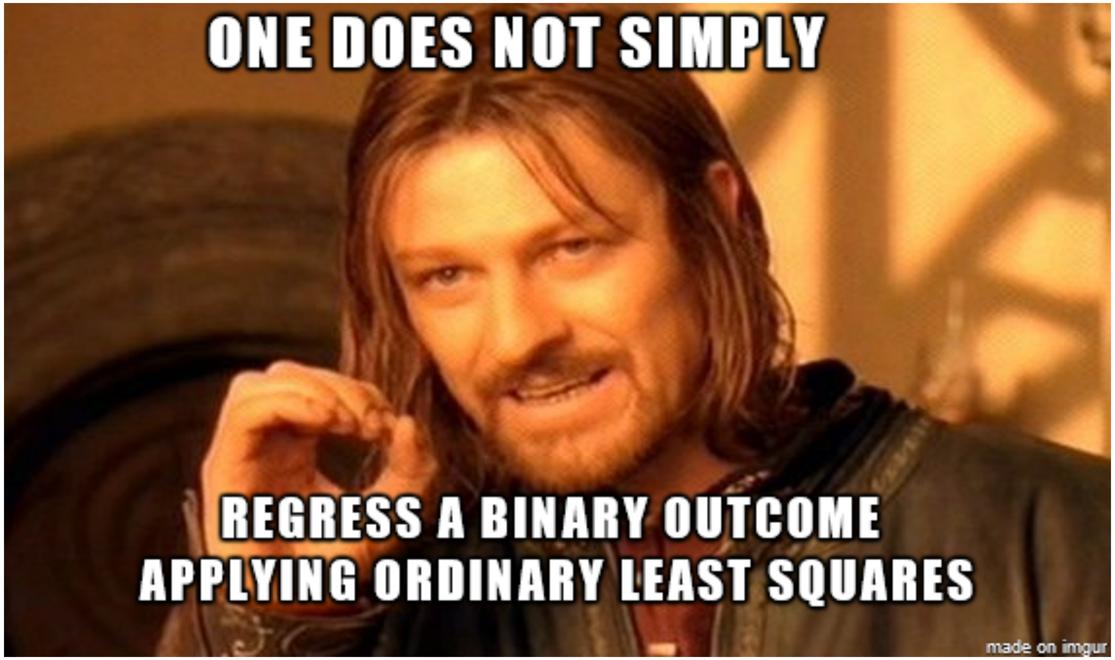
BST 210
Applied Regression Analysis



Lecture 10
Plan for Today

1. (brief) What if elements of LINE do not hold?
   - Which elements of LINE are non-negotiable and which can we relax a bit?
   - Do we still get good estimates of β and σ^2 in these scenarios? -> robust standard errors
   - A few extensions of linear regression in such scenarios
2. (less brief) Binary outcomes: Bin(n,p)
   - Are there extensions of models we already know, for modeling binary outcomes? Do they work?
   - Measures of effect: RD, RR, Odds and OR (CIs and SEs)
   - $\text{logit}(p) = \alpha + \beta x$; $p = ?$
   - Odds of success = e^{α} ; OR for exposed/unexposed = e^{β}
   - CIs, SEs, Hypothesis tests for logistic regression parameters

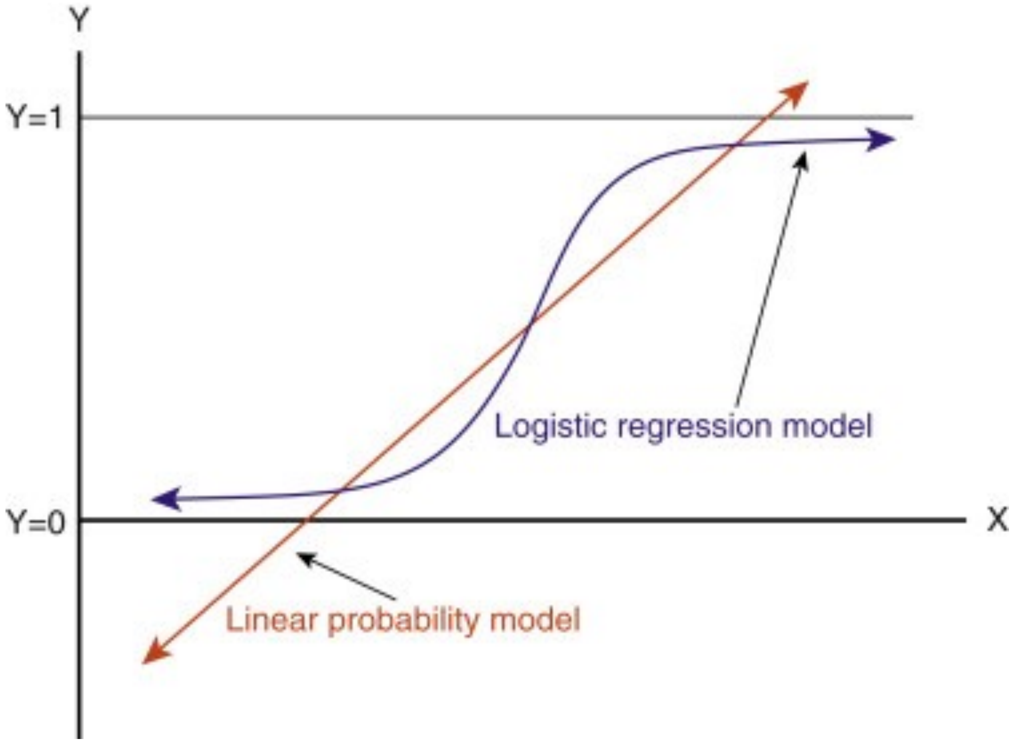
Linear Regression Assumptions and Extensions

- LINE for $E(Y_i) = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \ldots + \beta_p \cdot x_{ip}$
  - The mean of *Y* is a linear function of the covariates (quadratic terms, splines, etc.)
  - All responses are independent
  - The residuals are normally distributed
  - The residuals have equal variance (homoscedasticity)

What if LINE Not Satisfied?

- L: If linearity does not hold, that means our model for E(*Y*) is misspecified
- Our model E(Y) needs to be correct, or all bets are off
- There would be no reason to get consistent or unbiased estimates for β, even for large samples, with an incorrect specification of the regression model (our parameters and standard errors will be trying to estimate the wrong things!)

What if LINE Not Satisfied?



What if LINE Not Satisfied?

- L, N, E: In fact, as long as our model is correctly specified (Y is a linear function of the covariates), and our errors/residuals have mean 0 => our $\hat{\beta}$ OLS estimates will be unbiased for the true β.
- This holds whether or not we have independence, normality of residuals, or constant variance (—getting to this next!)
- Point: the main issues around linear regression assumptions tend to arise in our standard error estimates (—also next!)

What if LINE Not Satisfied?

- I: Independence is required in order for standard errors to be estimated appropriately – methods need to be extended if you have correlated, clustered, or longitudinal responses
- Generalized estimating equations (GEEs—topic in BST 226) for correlated responses with working independence would give β estimates after "fixing" the s.e. estimates so they are appropriate

What if LINE Not Satisfied?

- N and E: Here is where we can relax the assumptions a bit. In fact, provided that the residuals have mean zero and finite variance, we don't necessarily need them to be normally distributed or to have equal variance
- The residuals will have mean zero with an intercept in the model, one can show

What if LINE Not Satisfied?

- N and E: The scope of this part is for inference class, but one can show that $E(\hat{\beta}) = \beta$ for the least squares estimator if the residuals have mean 0, even when the residuals are not normal or homoscedastic
- Great for estimating β, but what about s.e. estimates, p-values, etc.? These inferences are indeed affected by the non-normality!

What if LINE Not Satisfied?

- N and E: Because N and E aren’t satisfied, we need to “adjust” the usual s.e. estimates of the β coefficients
- The resulting estimates are called *robust standard errors*, or Huber-White standard errors, or sandwich standard errors, or GEE standard errors (with independent responses)
- Given these, one can calculate CI’s and P-values

What if LINE Not Satisfied?

- N and E: These standard errors can be calculated easily in most packages
- And, since these s.e. estimates will get smaller as the sample size gets larger, we will still get consistency of the β estimates
- Note that with robust se estimation, we no longer get information about the Sum of Squares decomposition

What if LINE Not Satisfied?

- Some might recommend using the robust s.e. estimates all the time (rather than the usual s.e. estimates), but:
  - these don't lead to exact *t* tests or *F* tests as before
  - decrease efficiency /power if not assumptions do hold
  - they are asymptotic (large sample) s.e.'s, not "exact" for small samples, as the *t*-based CIs are for normal, homoscedastic errors
  - there are no $\hat{\sigma}^2$, adjusted R^2, MSE, AIC, BIC, leverages, studentized residuals, Cook's distances, etc.

Alternate Approaches

- *Weighted least squares* could be used when the residuals are normally distributed but have unequal variance
- But, you need to develop a way of setting $Var(Y_i)$ as a function of covariates

Alternate Approaches

- Various *robust regression methods* have been proposed that down-weight high residual observations, so as to give them less effect on the estimation of β

Bottom Line

- If you can verify the LINE assumptions through model assessment (histograms of residuals, QQ plots, etc.), then stick with “ordinary” least squares and multiple linear regression analysis
- It can possibly be useful to use robust methods when the N and E assumptions are not exactly satisfied, or to compare them at the end to your usual s.e. estimates

Many Extensions

- *Nonlinear least squares*, when the model for the mean is a nonlinear function of the β coefficients, perhaps still with normally distributed errors

Many Extensions: Longitudinal

- *Random or mixed effects models*, which allow some β coefficients to vary by subject when you have repeated measures on subjects
- *Generalized estimating equations* for adjusting the s.e. estimates, even if the model for the covariances are incorrect
- *Conditional models*, *time series models*, etc.
- (All a focus of BST 226 in Spring)

Next

- Start of logistic regression! – measures of association, odds ratios, logits, interpretation of logistic regression coefficients, examples!

Introduction to Logistic Regression

- We are often interested in public health data that is binary rather than continuous
  - Tumor metastasize/not
  - FEV threshold/not
  - Preeclampsia/not
  - Cancer remission/not

- Would LINE hold? Why or why not?
- What to do?

Introduction to Logistic Regression

- Let’s move now from the consideration of continuous outcomes using linear regression to

  → binary (0/1) outcomes using logistic regression

- Y = 0 (no event) or 1 (event)
- p = P(Y=1) = E(Y)

- Let’s review a little about binary outcomes and measures of association first

Review of Binomial Distribution

- How do we get to Binomial Y?
- Start with…***Bernoulli trials (Success/Failure)***… arising from the Bernoulli distribution, Bern(p):

$$P(X = x) = \begin{cases} p \ for\, x = 1 \\ 1 - p \ for\, x = 0 \end{cases}$$

- Also written $P(X=x) = p(1-p)$, $x = \{0,1\}$
- At the individual ('i') level, cancer/not, Afib/not

Review of Binomial Distribution

- Y ~ Bin(n,p) then arises as the sum of those n Bernoulli trials, where probability of success = p

  (pay attention to p as it has many uses!)

- Examples:
  - Total cases of asthma (asthma/not)
  - Total metastasized tumors (metastasized/not)
  - Total low weight births (low weight/not)

- Mathematically…

Review of Binomial Distribution

X ~ Bin (n, p):

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \qquad x = 0, 1, \ldots, n \quad 0 < p < 1$$

$\binom{n}{x}$ = number of ways of choosing x objects from a total of n without regard to order

$$\binom{n}{x} = \frac{n!}{(n-x)!x!} \qquad x! = 1(2)(3)\ldots(x) \qquad 0! \equiv 1$$

where

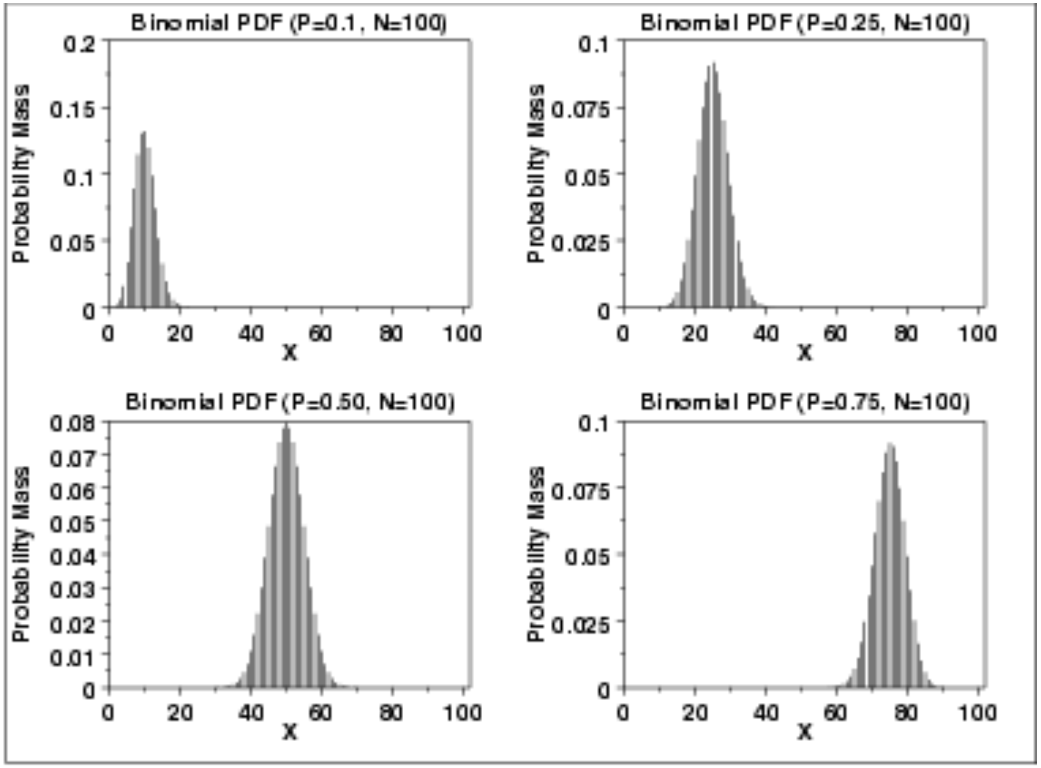
n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

q = 1 – *p* = the probability of getting a failure in one trial

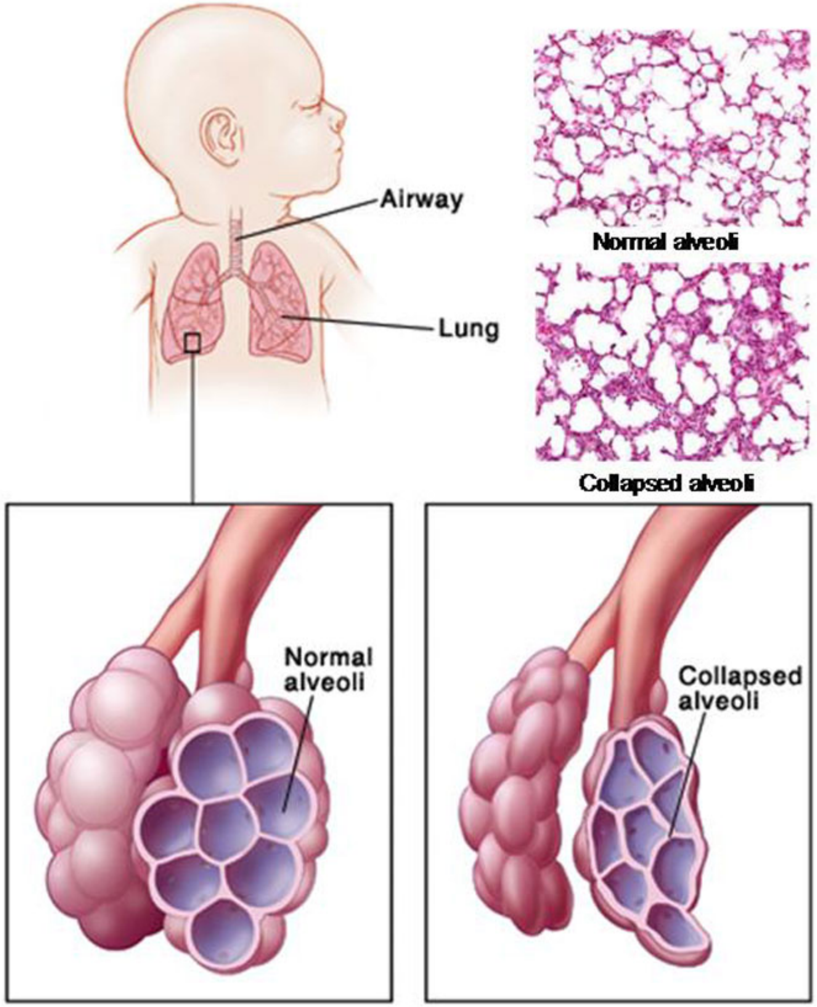
Review of Binomial Distribution



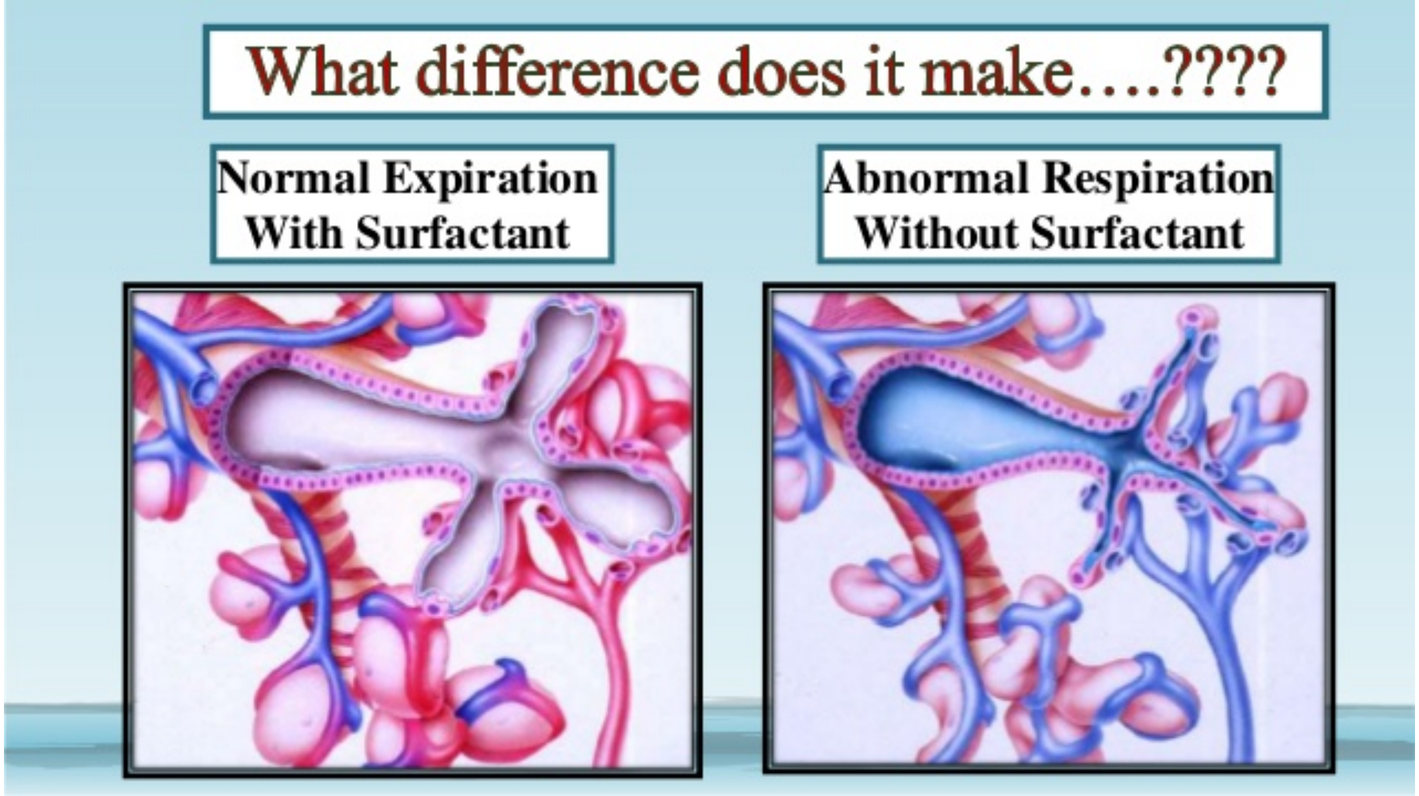
Example: Surfactant Use

- One of the leading causes of death for low birth weight babies is respiratory distress syndrome (RDS)
- During the period 1985-1990, use of surfactant (a compound that reduces the surface tension between two liquids) was very common as a treatment for RDS
- Surfactant is introduced intra-tracheally (i.e., through the windpipe) to infants with RDS

Example: Surfactant Use



Example: Surfactant Use


What difference does it make….????
Normal Expiration With Surfactant
Abnormal Respiration Without Surfactant

Example: Surfactant Use

- A study was performed comparing in-hospital mortality (0/1 variable) in low birth weight infants in 14 hospitals before and after the start of surfactant use

- Before: 3922 births with weight 500-1500 g, of which 960 died in the hospital (group 2)

- After: 1707 births with weight 500-1500 g, of which 335 died in the hospital (group 1)

$$\hat{p}_2 = 0.245, \hat{p}_1 = 0.196.$$

Example: Surfactant Use

		In-Hospital Mortality		
		Yes	No	Total
Surfactant Use	Yes	335 (19.6%)	1372	1707
	No	960 (24.5%)	2962	3922
	Total	1295 (23.0%)	4334	5629

Example: Surfactant Use

- Recall the *two-sample test for binomial proportions*
- Evaluate the null hypothesis that the proportion *p* of deaths is the same in the two groups

$$H_0 : p_1 = p_2$$

$$Z = \frac{(p_1 - p_2) - 0}{\sqrt{[p\,(1\text{-}p)\,(1/n_1 + 1/n_2)]}}$$

- This test statistic has a standard normal distribution if the null hypothesis is true (the Pearson χ^2 test could also have been used)

Example: Surfactant Use

- Conducting the test at the 0.05 level of significance,

  p = 1295/5629 = 0.230 and (1-p) = 0.770

- Therefore, Z = -3.94 and p < 0.001

- We reject H_0 in favor of the alternative at the 0.05 level of significance, and conclude that the two population proportions are not equal; the mortality rate is lower after the introduction of surfactants

- Do we have a magnitude of this association?

Measures of Effect

- We will look at 3 measures of effect as we motivate what underlies the need for and application of logistic regression!
- **Risk Difference, Risk Ratio, Odds Ratio**
- All utilize ‘proportion’ of success (or failures)

Measures of Effect: Risk

- The probability of having an outcome of interest is often called the *risk* of the outcome

- The probabilities of the outcome are actually conditional probabilities,

  p_1 = P(outcome | exposure) and

  p_2 = P(outcome | no exposure)

- Comparisons of these two conditional probabilities are called *measures of effect*

Measures of Effect: Risk Difference

- The *risk difference* $p_1 – p_2$ can be estimated by the difference in sample proportions (note q=1-p)

$$p_1 – p_2 \sim \; N(p_1 – p_2, p_1q_1/n_1 + p_2q_2/n_2)$$

- A 95% confidence interval for $p_1 – p_2$ is

$$p_1 – p_2 \pm 1.96 \sqrt{p_1q_1/n_1 + p_2q_2/n_2}$$

if $n_1 \; p_1 \; q_1 \geq 5$ and $n_2 \, p_2 \, q_2 \geq 5$

Measures of Effect: Risk Difference

RD = P(outcome | exposure) – P(outcome | no exposure)

RD = 0 No association between outcome and exposure

RD > 0 Positive association between outcome and exposure (probability of outcome is higher if exposed)

RD < 0 Inverse (or negative) association between outcome and exposure (probability of outcome is lower if exposed)

Example: Surfactant Use

- The estimated risk difference is

0.196 – 0.245 = -0.049

- A 95% confidence interval for the RD is given by

$$-0.049 \pm 1.96 \sqrt{[(.245)(.755)/3922 \ + \ (.196)(.804)/1707]}$$

or (-0.072, -0.025)

Example: Surfactant Use

- Note that the 95% confidence interval for $p_1 - p_2$ does not contain the value 0

- This tells us that we would reject H_0: $p_1 = p_2$ with p value < 0.05, and again conclude that the mortality rates are not the same

- Also note: The variance of the difference in sample proportions does not assume that $p_1 = p_2$ (unlike the assumption for the two-sample *z* test for binomial proportions)

Measures of Effect: Risk Ratio

- The *risk ratio* (RR) is a second way of estimating the magnitude of association

- It is defined as p_1/p_2 and can be estimated by

$$\hat{p}_1 / \hat{p}_2.$$

- It is a measure of relative rather than absolute risk

Measures of Effect: Risk Ratio

$$\mathbf{RR} = \frac{\mathbf{P(outcome \mid exposure)}}{\mathbf{P(outcome \mid no\ exposure)}}$$

RR = 1 No association between outcome and exposure

RR > 1 Positive association between outcome and exposure (probability of outcome is higher if exposed)

RR < 1 Inverse (or negative) association between outcome and exposure (probability of outcome is lower if exposed)

Measures of Effect: Risk Ratio

- The risk ratio is not as close to being normally distributed (it is positively skewed)

- The natural logarithm of the risk ratio approaches a normal distribution faster

- Therefore, to find a confidence interval for RR, we begin by finding a confidence interval for log(RR)

- Note that $\log(\mathrm{RR}) = \log(p_1/p_2) = \log(p_1) - \log(p_2)$

- The standard error estimates of $\log(p_1)$ and $\log(p_2)$ are found via the delta method (and we are using natural logs throughout, written as log or ln)

Risk Ratio

- Advantage: The risk ratio is very intuitive and easy to understand

- Disadvantage: RR is constrained by the probability of disease among the unexposed.

- If p_2 = 0.8, RR can be no larger than 1/0.8 = 1.25

- Therefore, it is difficult to combine RR estimates over low and high risk groups, since it is unlikely that they have a common relative risk

Odds

- The *odds* in favor of some outcome is defined as p/(1-p), where p = probability of the outcome

- If p = 0.5, then odds = 1 to 1

- If p = 0.8, then odds = 4 to 1

- If p = 0.2, then odds = 0.25 to 1

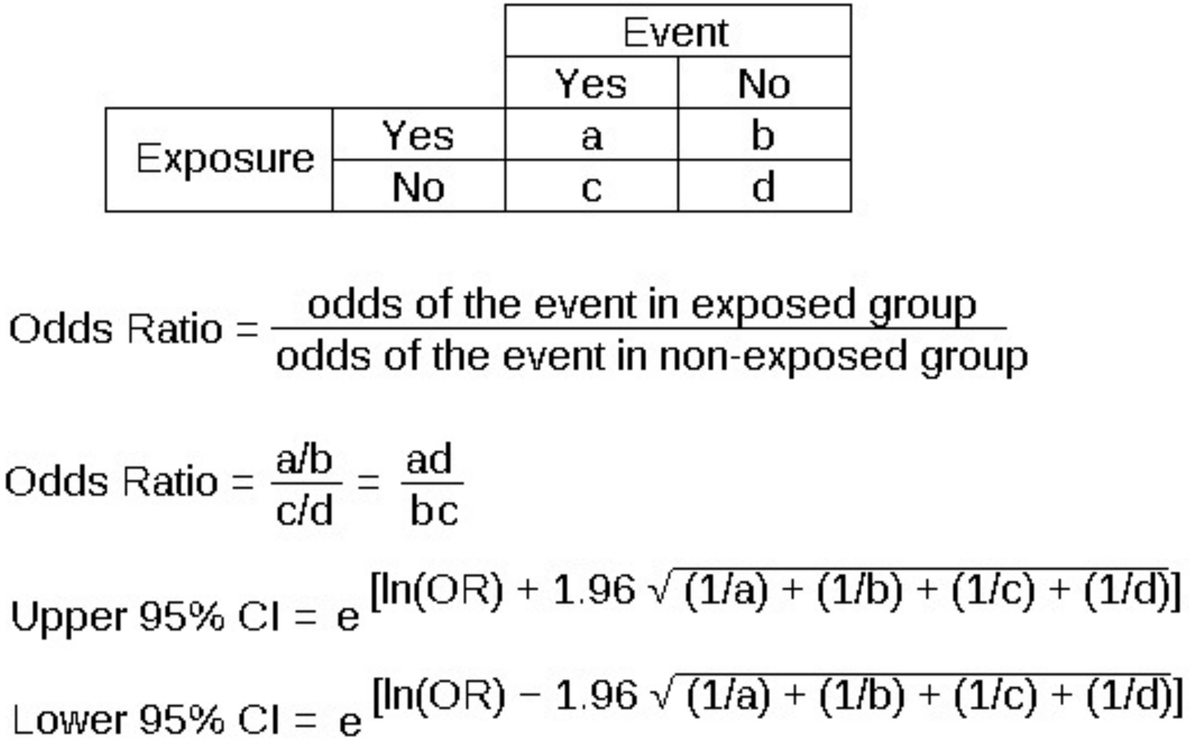
Odds Ratio

The *odds ratio* of an outcome for exposed versus unexposed subjects is defined as:

Odds in favor of outcome for exposed / odds in favor of outcome for unexposed

$$= \frac{p_1 / (1 - p_1)]}{p_2 / (1 - p_2)]}$$

Odds Ratio



		Event	
		Yes	No
Exposure	Yes	a	b
	No	c	d

$$\text{Odds Ratio} = \frac{\text{odds of the event in exposed group}}{\text{odds of the event in non-exposed group}}$$

$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

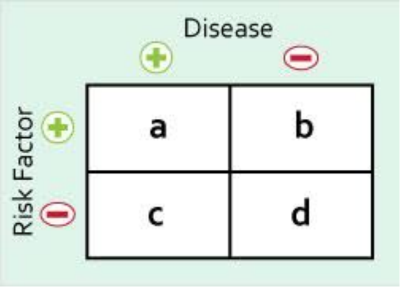
$$\text{Upper 95\% CI} = e^{[\ln(OR) + 1.96 \sqrt{(1/a) + (1/b) + (1/c) + (1/d)}]}$$

$$\text{Lower 95\% CI} = e^{[\ln(OR) - 1.96 \sqrt{(1/a) + (1/b) + (1/c) + (1/d)}]}$$

Odds Ratio

- If we have a 2 x 2 table with cell counts a, b, c and d, then the estimated OR is

$$(a/b) / (c/d) = ad / bc$$



Risk Factor	Disease ⊕	Disease ⊖
⊕	a	b
⊖	c	d

- The OR ranges from $+\infty$ when $p_1 = 1$ and $p_2 < 1$, to 0 when $p_1 = 0$ and $p_2 > 0$

- Therefore, it is more plausible that low and high risk populations have the same OR

Odds Ratio

- The odds ratio may be less intuitive than the relative risk

- However, when the probability of disease is small, OR $\cong$ RR since $1 - p_1$ and $1 - p_2$ are both close to 1

- For example, if $p = 0.02$,

  then $p/(1 - p) = 0.02/0.98 = 0.0204 \approx 0.02$

- Also, there are some research designs where the risk ratio cannot be estimated

Using Software for RD, RR, and OR

```
. csi 335 960 1372 2962, or woolf

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |       335         960  |       1295
        Noncases |      1372        2962  |       4334
-----------------+------------------------+------------
           Total |      1707        3922  |       5629
                 |                        |
            Risk |  .1962507    .2447731  |   .2300586
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0485223       |   -.0716748   -.0253699
      Risk ratio |          .801766       |     .717798    .8955566
 Prev. frac. ex. |          .198234       |    .1044434     .282202
 Prev. frac. pop |         .0601147       |
      Odds ratio |         .7533634       |    .6550237     .866467
(Woolf)
                 +-------------------------------------------------
                               chi2(1) =    15.81  Pr>chi2 = 0.0001
```

This leads us to:

Regression Models for Binary Outcomes

- Suppose that the outcome variable *Y* can only assume two possible values (e.g., success and failure, coded as 1 and 0)

$$p = \mathrm{P}(Y = 1) = \mathrm{P}(\text{success})$$

- We would like to be able to estimate the probability *p* associated with a particular value of an explanatory variable *X*

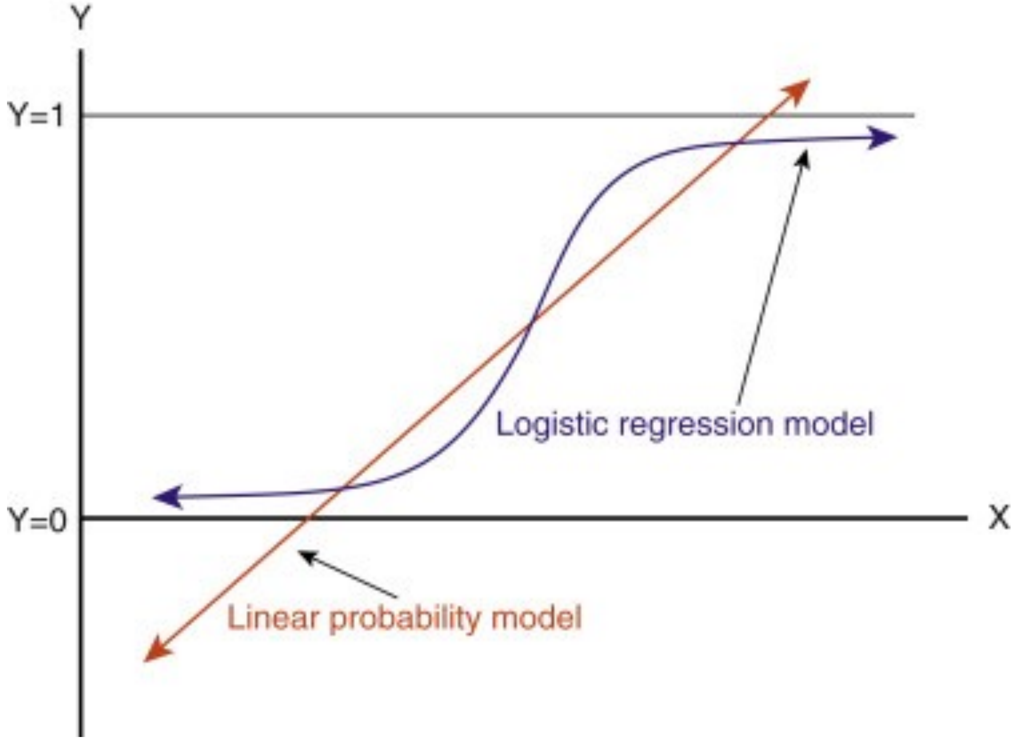
Regression Models for Binary Outcomes

- Given a covariate (risk factor or exposure or explanatory variable) X and probability of outcome p, we could try to fit a linear model $p = \alpha + \beta X$
- However, the predicted probabilities from this model could be < 0 or > 1, which is impossible
- Furthermore, for a binary outcome having probability p, the variance is $p\,(1 – p)$, which is not constant if p changes as a function of a covariate X

Regression Models for Binary Outcomes

- There are similar concerns for an exponential model p = exp($\alpha + \beta\, X$) = $e^{\alpha + \beta X}$
- Predicted probabilities from this model can be > 1
- How can we guarantee that our probabilities will fall between 0 and 1?

Regression Models for Binary Outcomes



Logits

- **Define logit(*p*) = log[*p*/(1 – *p*)] = log(odds)**

p = 0.2, logit(0.2) = log(0.2/0.8) = log(1/4) = -1.39

p = 0.5, logit(0.5)= log(0.5/0.5) = log(1) = 0

p = 0.8, logit(0.8) = log(0.8/0.2) = log(4) = 1.39

p = 0, logit(0) = log(0) = $-\infty$

p = 1, logit(1) = log(∞) = ∞

Logistic Regression Model

- We could use the model defined as

$$\log[p/(1 - p)] = \alpha + \beta X$$

- We are fitting a linear model on the logit scale

- We assume that the relationship between $\log[p/(1 - p)]$ and X is linear
- What about p?

Logistic Function

- Solving for *p*, we obtain:

$$\boldsymbol{p = \exp(\alpha + \beta X) / [1 + \exp(\alpha + \beta X)]}$$

- In this model, estimated probabilities are restricted to falling between 0 and 1

- This expression called a *logistic function*

Logistic Regression Model

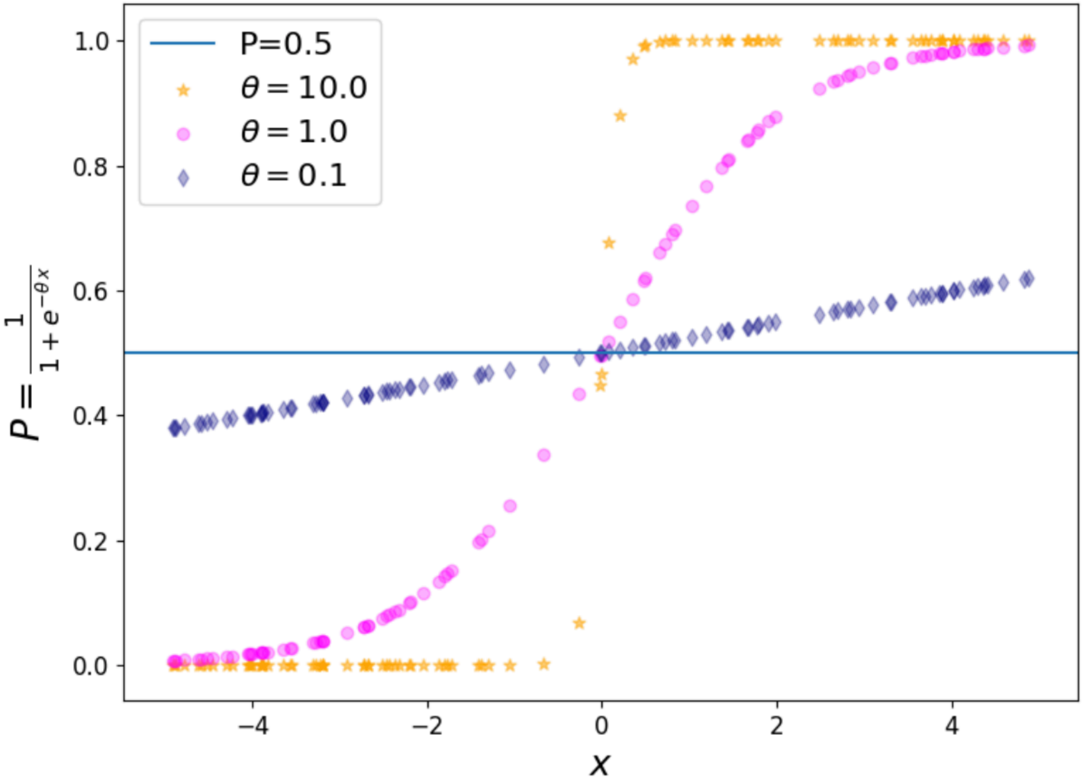


Figure 2: Probability vs independent variable **x**; resembles sigmoid function plot.

Logistic Regression

- Let the subscript *i* represent the *i*th subject in a sample

- Let *X* be a binary covariate such that x_i = 1 if the subject is exposed and x_i = 0 if unexposed

- p_i = probability of disease for the *i*th subject

- We fit the logistic regression model

$$\textbf{logit}(\boldsymbol{p}_\mathrm{i}) = \textbf{log}[\boldsymbol{p}_\mathrm{i}/(1 - \boldsymbol{p}_\mathrm{i})] = \alpha + \beta\, \boldsymbol{x}_\mathrm{i}$$

Logistic Regression

- The constant α is the intercept of the regression model

- If $x_i = 0$, then

$$\begin{aligned} \text{logit}(p_i) &= \log[p_i/(1 - p_i)] \\ &= \log(\text{odds of success}) \\ &= \alpha \end{aligned}$$

- If $x_i = 0$, the odds of success are $\exp(\alpha)$

Logistic Regression

- Compare logit(p_i) for an exposed subject versus an unexposed subject

- Exposed: $\log[p_i/(1 - p_i)] = \alpha + \beta(1)$
  $= \alpha + \beta$

- Unexposed: $\log[p_i/(1 - p_i)] = \alpha + \beta(0)$
  $= \alpha$

Logistic Regression

$$
\begin{aligned}
\beta &= (\alpha + \beta) - \alpha \\
&= \text{logit}(p_\text{i} \mid x_\text{i} = 1) - \text{logit}(p_\text{i} \mid x_\text{i} = 0) \\
&= \log(\text{odds} \mid x_\text{i} = 1) - \log(\text{odds} \mid x_\text{i} = 0) \\
&= \log([\text{odds} \mid x_\text{i} = 1] / [\text{odds} \mid x_\text{i} = 0]) \\
&= \log(\text{odds ratio})
\end{aligned}
$$

Thus the odds ratio for exposed versus unexposed subjects is exp(β)

Coming Up

- More on logistic regression – confounding, effect modification, model building, interpretation