BST 210
Applied Regression Analysis

Lecture 13

Plan for Today

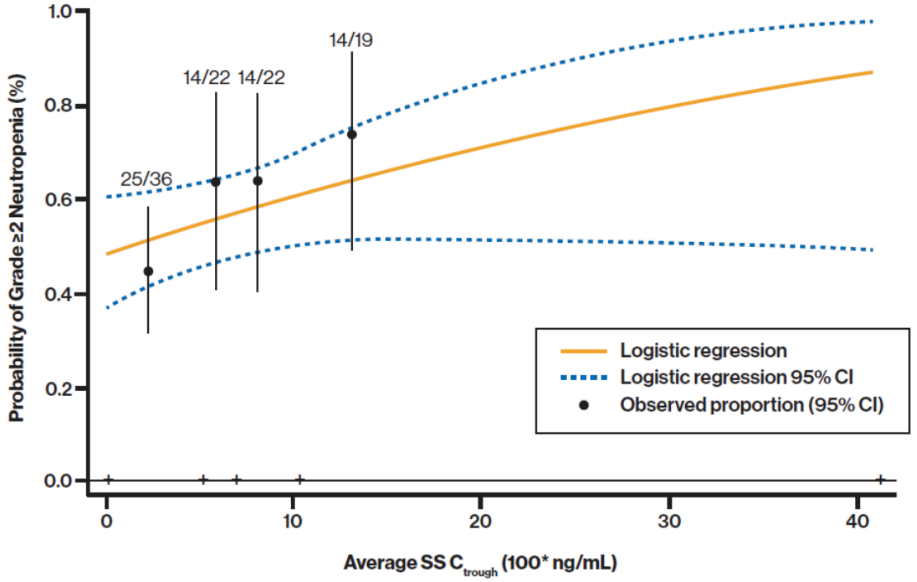
- 5 examples of logistic regression out in the world!
- Confounding and Effect Modification continued…
  - Examples and interpretations
- Hypothesis testing in Logistic Regression
  - Comparison to Linear Regression
  - Wald test (vs t-test)
  - LRT (likelihood ratio test) (vs F test)

* Note: 'Framework for Analysis' will be back! (being re-worked)

Example 1

Updated Results From MONALEESA-2, a Phase 3 Trial of First-line Ribociclib + Letrozole in Hormone Receptor-Positive (HR+), HER2-Negative (HER2–), Advanced Breast Cancer (ABC)

Logistic Regression of Grade ≥2 Neutropenia vs Ribociclib Exposure – Pooled Analysis



CI, confidence interval; C_{trough}, trough plasma concentration; SS, steady state.
Pooled analysis of 276 patients with cancer treated with ribociclib 50–1200 mg/day (3-weeks-on/1-week-off) in the following clinical trials: NCT01898845, NCT01237236, NCT01872260, and NCT01958021 (MONALEESA-2).

Janni W, et al. *J Clin Oncol*. 2017;35(suppl): Abstract 1047.

Example 2

Multiple Logistic Regression Analysis of Risk Factors Associated with Denture Plaque and Staining in Chinese Removable Denture Wearers over 40 Years Old in Xi'an – a Cross-Sectional Study

Variable	*P*	OR	95% CI
Duration of denture use (years)	<0.001		
≤0.5	–	1.000	–
0.6–2	0.110	2.087	0.846–5.152
2.1–5	0.001	4.155	1.801–9.590
>5	<0.001	7.238	3.275–15.995
Main cleaning method	0.009		
Running water	0.010	7.081	1.590–31.528
Brushing	0.005	3.567	1.459–8.720
Chemical cleanser	–	1.000	–

Denture plaque scores were dichotomized using the median (1.67) as a cutoff value (≤1.67, >1.67). Independent variables included duration of denture use, denture wear status, and cleaning and overnight storage methods. Omnibus tests of model coefficients indicated that the χ^2 value of the logistic regression model was 40.129 and the *P* value was <0.001.

doi:10.1371/journal.pone.0087749.t003

Example 3

CCO CHINESE CLINICAL ONCOLOGY
AN OPEN ACCESS JOURNAL FOR CLINICAL CANCER RESEARCH WITH CURRENT AND PRACTICAL INFORMATION
Chinese Clinical Oncology

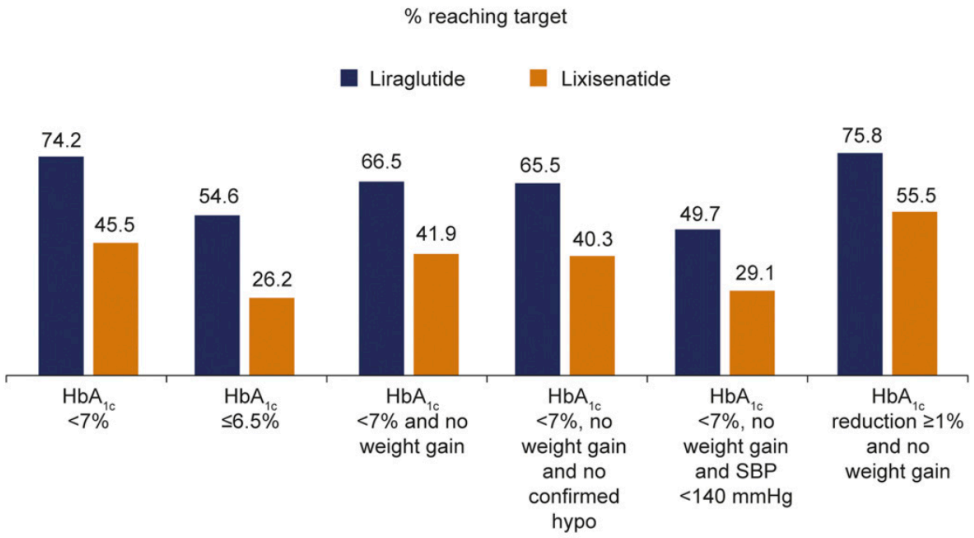
Table 4 Main methods of analysis for phase III cancer clinical trials

Purpose of analysis	Nature of endpoint		
	Normal (e.g., white blood cell counts)	Binary (e.g., tumor response)	Time-dependent (e.g., survival)
Estimation	Mean (95% CI) and quantiles	Proportion (95% CI)	Median (95% CI) and Kaplan-Meier curves
Hypothesis test (unadjusted)	*t*-test or wilcoxon test	χ^2 test or fisher exact test	Logrank test
Hypothesis test (adjusted for covariates)	Analysis of variance	Mantel-Haenszel χ^2 test	Stratified logrank test
Regression analysis (with covariates)	Linear regression model	Logistic regression model	Cox regression model

CI, confidence interval.

Example 4

American Diabetes Association. Diabetes Care



OR lira-lixi	4.2	3.7	3.1	3.2	2.7	2.7
95% CI	[2.6 ; 6.7]	[2.3 ; 5.8]	[2.0 ; 4.8]	[2.0 ; 5.0]	[1.7 ; 4.3]	[1.7 ; 4.3]
P value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Figure 2—Percentage of patients reaching targets. Patients meeting targets (%) and estimated ORs based on a logistic regression model, with treatment and country as fixed factors and the HbA_{1c} value at baseline as a covariate. Hypo, hypoglycemia; Lira, liraglutide; Lixi, lixisenatide.

Example 5

The NEW ENGLAND JOURNAL of MEDICINE

A Placebo-Controlled Trial of Antibiotics for Smaller Skin Abscesses

Table 4. Logistic-Regression Model of Cure Rates among Children and Adults.

Variable and Population	P Value in the Logistic-Regression Model*		
	TMP-SMX vs. Clindamycin	Placebo vs. Clindamycin	Placebo vs. TMP-SMX
Study group			
Intention-to-treat population	0.37	<0.001	0.003
Population that could be evaluated	0.17	<0.001	<0.001
Age group			
Intention-to-treat population	0.11	0.11	0.98
Population that could be evaluated	0.04	0.03	0.87
Interaction			
Intention-to-treat population	0.17	0.09	0.83
Population that could be evaluated	0.06	0.06	0.74

* The P values refer to the results of a logistic-regression model incorporating study group (clindamycin vs. TMP-SMX) and age group (children vs. adults). After controlling for the effect of age, the differences between placebo and clindamycin and between placebo and TMP-SMX were significant in both the intention-to-treat population and the population that could be evaluated (P<0.001). The cure rates for clindamycin in the population that could be evaluated were significantly higher among children than among adults (among children, P=0.04 for TMP-SMX vs. clindamycin and P=0.03 for placebo vs. clindamycin). None of the interaction terms for the logistic-regression models were significant, indicating that the differences in cure rates between children and adults were not significant in each respective study-group comparison (P>0.05 for the intention-to-treat population and the population that could be evaluated).

Recall: Confounding

- Is a *bias* in your estimate of the exposure-disease association due to a third variable *X*, the confounding variable
- The crude or unadjusted OR may be inappropriate and the adjusted OR should be used instead
- Can sometimes be controlled for with a proper analysis or avoided by design
- Is a bias, and worth avoiding!

Recall: Effect Modification

- Is a *change* in the exposure-disease association across levels of a third variable *X*, the effect modifier
- Either a crude or adjusted OR may be inappropriate, and separate ORs should be described according to the level of *X*
- Is a property of the true exposure-disease association, and exists (or doesn't exist) independent of your study or study design
- Is an interesting phenomenon, worthy of description

Confounding vs. Effect Modification

- We use multivariable logistic regression methods to assess the possible effects of an additional variable on the primary exposure-disease association of interest

- This additional variable (or variables) may serve as confounders or effect modifiers. Usually we start by assessing confounding, but if there is significant effect modification we want to account for that in our inferences

More on Effect Modification (EM)

- If effect modification (or an interaction) exists between an exposure and a categorical effect modifier, then it may be either a qualitative or a quantitative interaction

- *Qualitative interaction*: $OR_i > 1$ for some categories, $OR_i < 1$ for other categories

- *Quantitative interaction*: Either $OR_i > 1$ for all categories, or $OR_i < 1$ for all categories

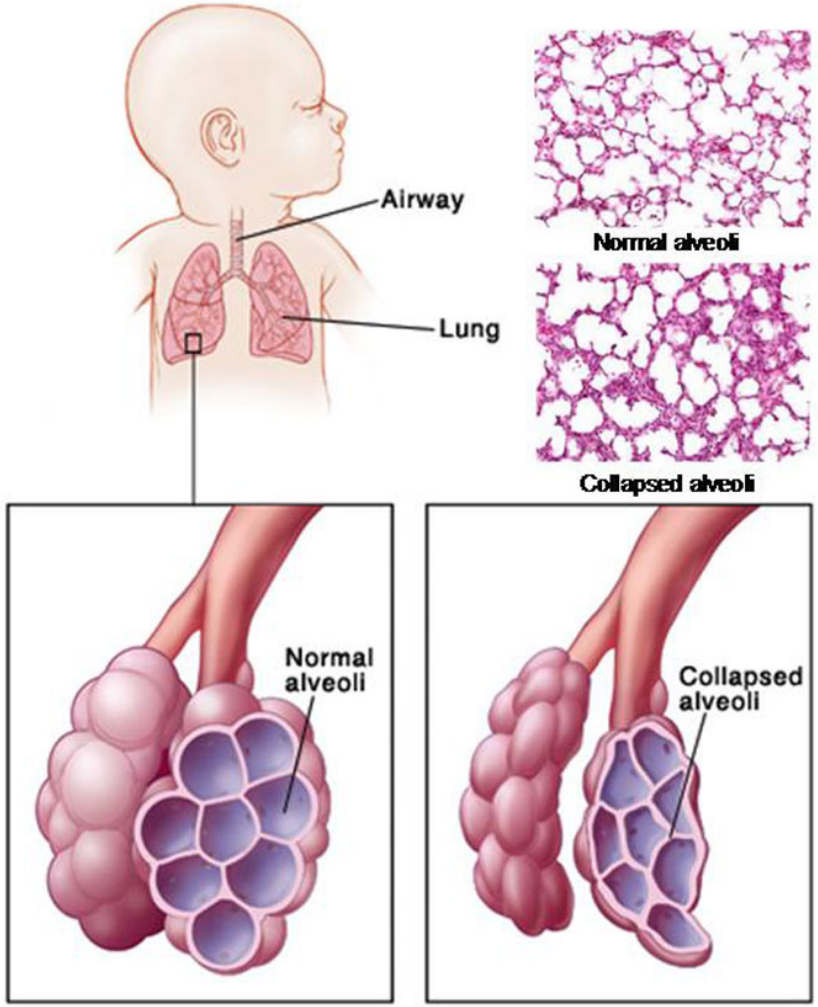
(We often call these categories *strata*)

EM: Qualitative versus Quantitative

- For qualitative interactions, results should be presented for each stratum separately without pooling or collapsing

- For quantitative interaction, it may still might make sense to estimate a common odds ratio as an “average effect”, in addition to calculating stratum-specific odds ratios, but the stratum-specific odds ratios are probably more appropriate

Recall Example: Surfactant Use



Recall 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)

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

Recall Example: Surfactant Use

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

Remember: **OR** = (335)(2962)/(960)(1372) = 0.75

Recall Covariate: Birth weight

- We've seen in a prior class that birth weight is a potential covariate of interest, and is divided into the following 4 categories:

500-749 g	(group 1)
750-999 g	(group 2)
1000-1249 g	(group 3)
1250-1500 g	(group 4)

Recall Covariate: Birth weight

- And for modeling we define

  x_2 = 1 if birth weight = group 2, 0 otherwise

  x_3 = 1 if birth weight = group 3, 0 otherwise

  x_4 = 1 if birth weight = group 4, 0 otherwise

- We have already run logistic regression models with birth weight as a covariate (categorical and ordinal 'continuous')

$$\log[p/(1-p)] = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Focus first on: Controlling for Confounding

- We next denote the full model with surfactant and birthweight:

$$\text{logit}(p_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

where

x_{1i} = surfactant use (1 = yes, 0 = no)

x_{2i} = 1 if birth weight = 750-999 g, 0 otherwise

x_{3i} = 1 if birth weight = 1000-1249 g, 0 otherwise

x_{4i} = 1 if birth weight = 1250-1500g, 0 otherwise

Controlling for Confounding

- How do we interpret the parameters of this model?
- Consider two individuals A and B where subject A was born in the time period when surfactant was used ($x_1 = 1$) and subject B was born before surfactant was used ($x_1 = 0$)
- They are both in the same birth weight group (e.g., group 3 = 1000-1249 g, $x_3 = 1$)

Controlling for Confounding

- It follows that:

$\text{logit}(p_A) = \alpha + \beta_1 + \beta_3$
$\text{logit}(p_B) = \alpha + \beta_3$

$\text{logit}(p_A) - \text{logit}(p_B) = \log[p_A/(1 - p_A) \,/\, p_B/(1 - p_B)] = \beta_1$

$e^{\beta 1} = OR_{A \text{ vs } B}$

- This is interpreted as the odds ratio for mortality associated with surfactant use, adjusted for birth weight group

Controlling for Confounding

- Similarly, consider two infants C and D where subject C is in birth weight group 3 ($x_3 = 1$), subject D is in birth weight group 1 ($x_2 = x_3 = x_4 = 0$), and both infants were born after surfactant use was introduced

$$\text{logit}(p_C) = \alpha + \beta_1 + \beta_3$$

$$\text{logit}(p_D) = \alpha + \beta_1$$

Controlling for Confounding

$$\text{logit}(p_C) - \text{logit}(p_D)$$
$$= \log[p_C/(1-p_C) \,/\, p_D/(1-p_D)]$$
$$= \beta_3$$

$$e^{\beta 3} = OR_{C \text{ vs } D}$$

- This is interpreted as the effect of birth weight 1000-1249 g versus 500-749 g on mortality, adjusting for differences in surfactant use

Controlling for Confounding

- Let's assess whether birth weight confounds the association between surfactant-use and death:
- First consider classical definition of confounder, then model surfactant-use alone, followed by surfactant-use with birth weight, and use 10% rule

```
. logit death surfactant

Iteration 0:   log likelihood = -3035.9852
Iteration 1:   log likelihood = -3027.9267
Iteration 2:   log likelihood = -3027.9091
Iteration 3:   log likelihood = -3027.9091

Logistic regression                                     Number of obs   =       5629
                                                        LR chi2(1)      =      16.15
                                                        Prob > chi2     =     0.0001
Log likelihood = -3027.9091                             Pseudo R2       =     0.0027

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |  -.2832076   .0713668    -3.97   0.000    -.4230838   -.1433313
       _cons |  -1.126687   .0371386   -30.34   0.000    -1.199477   -1.053896
------------------------------------------------------------------------------
```

Controlling for Confounding

```
. logit death surfactant bwt2 bwt3 bwt4

Iteration 0:   log likelihood = -3035.9852
Iteration 1:   log likelihood = -2514.6216
Iteration 2:   log likelihood = -2469.1618
Iteration 3:   log likelihood = -2468.5176
Iteration 4:   log likelihood =  -2468.517
Iteration 5:   log likelihood =  -2468.517

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(4)      =    1134.94
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2468.517                       Pseudo R2       =     0.1869

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |  -.3558706   .0801226    -4.44   0.000    -.5129081   -.1988331
        bwt2 |  -1.472116   .0885312   -16.63   0.000    -1.645634   -1.298598
        bwt3 |   -2.34365    .100741   -23.26   0.000    -2.541099   -2.146201
        bwt4 |  -3.006796   .1124737   -26.73   0.000    -3.227241   -2.786352
       _cons |   .5017456   .0665808     7.54   0.000     .3712495    .6322416
------------------------------------------------------------------------------

. display 100 * (-.2832076 - (-.3558706)) / (-.2832076)
-25.65715

. * Here the % change in the beta coefficients is
. * 100 * (crude beta - adjusted beta) / crude beta
. * As there is a 26% change in the beta coefficient (going from crude to
. * adjusted) we might say that we do have evidence for confounding, and
. * hence would prefer the adjusted OR estimate, as here it could be that
. * birth weight is associated with both surfactant as well as mortality
. * outcome, but birth weight is not on the causal pathway from surfactant
. * to mortality
```

Controlling for Confounding

- Thus we see that there is still a significant effect of surfactant on mortality after controlling for birth weight group (OR = 0.70, 95% CI [0.60, 0.82], $p < 0.001$)

- (This is slightly stronger than for the crude analyses from previous class: OR = 0.75, 95% CI [0.66, 0.87], $p < 0.001$)

- After controlling for birth weight, the odds of in-hospital mortality are 0.70 times lower for infants treated with surfactant compared to infants not treated with surfactant

Controlling for Confounding

- There are also significant effects of each birth weight category versus the reference group (all $p < 0.001$)

  Group 2 versus 1: OR = 0.23 (0.19, 0.27)

  Group 3 versus 1: OR = 0.10 (0.08, 0.12)

  Group 4 versus 1: OR = 0.05 (0.04, 0.06)

- They are similar to the unadjusted effects. What do you notice here?

- *(Note: We can also compare 'nested' models using a Likelihood Ratio Test in order to assess whether or not a covariate is significant – next! – but first we progress from assessment for confounding into that for effect modification.)*

Checking for Effect Modification

- We note in the surfactant-use data that the effect of birth weight is statistically significant, AND the odds of death decreases as birth weight increases

- In this analysis, one assumption that is made is that the effect of surfactant-use is the *same* for all birth weight groups

- However, looking at the original data, it seems possible that surfactant contributes markedly less effect in preventing death for lower versus higher birth weight categories (.23 versus .10, .05)

- This could be possible effect modification (or an interaction); birth weight might be *modifying* the relationship between surfactant-use and death

Checking for Effect Modification

- To look for effect modification, we can fit the model:

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4$$

- The terms x_1, x_2, x_3, x_4 are called *main effects*
- The product terms $x_1 x_2$, $x_1 x_3$, and $x_1 x_4$ are called *interaction terms*
- *(It's easier to see with a single binary covariate for sure!—more examples like that to come…)*

Checking for Effect Modification

- Main effects have a different interpretation if there is an interaction in the model

- The effect of surfactant-use on infant death is quantified by the coefficients of x_1:

$$\beta_1 + \beta_5 x_2 + \beta_6 x_3 + \beta_7 x_4$$

- If $\beta_5 \neq 0$, $\beta_6 \neq 0$, $\beta_7 \neq 0$ then the effect of surfactant-use varies with birth weight

Checking for Effect Modification

- For birth weight < 750 g, say (and we could consider various levels), β_1 = log(odds of death) and exp(β_1) is the relevant odds ratio relating surfactant-use to death, adjusting for birth weight

- For birth weight $\geqq$ 750 g, say, ($\beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7$) = log(odds of death) and exp($\beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7$) is the odds ratio relating surfactant-use to death adjusting for birth weight

- Let's run the full multiple logistic regression model (which includes interaction terms) above! →

Effect Modification

```
. * Finally, let's consider possible effect modification between surfactant
. * and birth weight category. One easy way to include both main effects plus
. * their interaction in Stata is to use # or ## options
```

```
. logistic death i.surfactant##i.birthwt

Logistic regression                                     Number of obs   =       5629
                                                        LR chi2(7)      =    1141.37
                                                        Prob > chi2     =     0.0000
Log likelihood = -2465.2995                             Pseudo R2       =     0.1880

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.surfactant |   .7003971   .0932387    -2.68   0.007     .5395481     .909198
             |
     birthwt |
     750-999 |   .2408591   .0252328   -13.59   0.000     .1961507    .2957579
    1000-1249|   .0971421   .0113736   -19.91   0.000     .0772231     .122199
    1250-1500|   .0435102   .0059352   -22.98   0.000     .0333027    .0568463
             |
surfactant#birthwt |
   1#750-999 |   .8331254   .1644391    -0.92   0.355     .5658528    1.226641
  1#1000-1249|   .9491598   .2196654    -0.23   0.822     .6030379    1.493943
  1#1250-1500|   1.564313   .3759617     1.86   0.063     .9766723    2.505524
             |
       _cons |   1.651724   .1228949     6.74   0.000     1.427594    1.911042
------------------------------------------------------------------------------
```

Effect Modification

- The coefficient of the interaction term(s) are not significantly different from 0, ($p \geq .05$)
- This suggests that there is no statistically significant effect modification by birth weight (as defined by these categories)
- The odds ratio of death for the surfactant-use group versus the non-surfactant-use group is not significantly different depending on an infant's birth weight
- We could also conduct a stratified analysis, but the multiple logistic regression model allows for inclusion of other potential covariates (which could be potential confounders or effect modifiers!)

Next: Hypothesis Testing in Logistic Regression

- We've now developed the logistic regression model out of need for modeling binary outcome data, and we've learned in depth how to interpret its parameters across various examples.

- We now want to better understand another pillar of the logistic regression framework—that which allows us to assess significance of predictor variables, compare nested models, and draw statistical inference from the results of our model selections and diagnostics.

→ ***Hypothesis testing!***

- Just as one would imagine, the concept of hypothesis testing in the logistic regression setting is quite similar to that in the linear regression context—with some differences.

Hypothesis Testing in Logistic Regression

Goal is same in both settings: we want to draw inference about our predictors X and their relationship with Y!

Linear Regression (slope):

- **t-tests:**

Formal Hypothesis: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

Test Statistic: $t = \frac{\hat{\beta}_j - 0}{\text{s.e.}(\hat{\beta}_j)} \sim t_{n-(p+1)}$

Confidence Interval: $\hat{\beta}_j \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j)$

- **F-tests:**

Formal Hypothesis: $H_0 : \beta_i = \beta_j = \beta_k = 0$ versus H_1 : at least one of β_i, β_j and β_k is not 0

or

Formal Hypothesis: $\left(H_0 : \text{the reduced model is sufficient}\right)$ versus $\left(H_1 : \text{the full model is preferred}\right)$

Test Statistic: $F = \frac{(SSE_{reduced} - SSE_{full})/3}{SSE_{full}/(n-(p+1))} \sim F_{3,n-p-1}$

Hypothesis Testing in Logistic Regression

Similarly,

Logistic Regression (OR):

- **Wald tests:**

Formal Hypothesis:	$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$	
Test Statistic:	$Z = \frac{\hat{\beta}_j - 0}{\text{s.d.}(\hat{\beta}_j)} \sim N(0,1)$	
Confidence Interval:	$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.d.}(\hat{\beta}_j)$	(log odds ratio scale)
	$\exp\{\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.d.}(\hat{\beta}_j)\}$	(odds ratio scale)

- **Likelihood Ratio Tests:**

The number of predictors in the full model is greater than the number in the reduced model, $p > q$.

Formal Hypothesis:	$(H_0$: the reduced model is sufficient$)$ versus $(H_1$: the full model is preferred$)$
Test Statistic:	$-2 \cdot \log\left(\frac{L_{reduced}}{L_{full}}\right) = -2 \cdot \log(L_{reduced}) + 2 \cdot \log(L_{full}) \sim \chi^2_{p-q}$

Hypothesis Testing for p

Let's back up -Simplest hypothesis test (for p) in this setting:

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

$$H_0 : p_1 = p_2$$

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

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

Hypothesis Testing for p

- (From previous class meeting) Conducting the test at the 0.05 level of significance,

  p = 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

- *This approach is limited—we want magnitude of the association, and ability to consider other factors in the story*

Hypothesis Testing for OR

Now involving the OR:

- We wish to test the null hypothesis

H_0: $\beta = 0$ versus the alternative hypothesis

H_1: $\beta \neq 0$

- The Wald test statistic

$$Z = (\hat{\beta} - 0)/\text{se}(\hat{\beta}) = \hat{\beta}\ /\text{se}(\hat{\beta})$$

has (approximately) a standard normal distribution if the null hypothesis is true

- In the surfactant example, we find $Z = -3.97$, $P < 0.001$, thus the odds of mortality are lower for infants treated with surfactants

Hypothesis Testing for OR

Confidence Intervals via Woolf Method

The point estimate of $OR = \hat{OR} = ad/(bc)$.

As was true for RR, $\ln(\hat{OR})$ is more closely normally distributed than $\hat{OR}$ itself.

It can be shown using the delta method that :

$$\text{var}[\ln(\hat{OR})] \approx 1/a + 1/b + 1/c + 1/d.$$

Hypothesis Testing for OR

Hence, $\ln(\hat{OR}) \approx N[\ln(OR), 1/a + 1/b + 1/c + 1/d]$.

Thus, an approximate $100\% \times (1-\alpha)$ CI for $\ln(OR)$ is given by :

$$\ln(\hat{OR}) \pm z_{1-\alpha/2}\sqrt{1/a + 1/b + 1/c + 1/d} = (c_1, c_2).$$

The corresponding $100\% \times (1-\alpha)$ CI for OR is given by $[\exp(c_1), \exp(c_2)]$.

As was true for RR, the CI for OR is not symmetric about $\hat{OR}$.

Hypothesis Testing for OR

For the surfactant data, we have $\hat{p}_1 = 0.196, \hat{p}_2 = 0.245.$

Hence,

$$\hat{OR} = \frac{0.196/0.804}{0.245/0.755} = 0.753, \text{ and } \ln(\hat{OR}) = -0.283.$$

using Woolf's method, we note that

a = 335, b = 1707-335 = 1372, c = 960 and

d = 3922-960 = 2962.

Hence, $\text{var}[\ln(\hat{OR})] = 1/335 + 1/1372 + 1/960 + 1/2962 = 5.09 \times 10^{-3}.$

$se[\ln(\hat{OR})] = 0.071.$

Hypothesis Testing for OR

A 95% CI for ln(OR) is thus given by:

$$-0.283 \pm 1.96(0.071) = (-0.423,\ -0.143) = (c_1, c_2).$$

The corresponding 95% CI for OR is:

$$\left[\exp(-0.423),\ \exp(-0.143)\right] = (0.66,\ 0.87).$$

Thus, in summary, the estimated OR

$= 0.75$ (95% CI = 0.66, 0.87).

Hypothesis Testing for Beta

- The coefficient $\hat{\beta}$ = ln(OR) is a maximum likelihood estimate (MLE), which has an approximate normal distribution

- A 95% confidence interval for β is:

$$(\hat{\beta} - 1.96 \text{ s.e.}(\hat{\beta}),\ \hat{\beta} + 1.96 \text{ s.e.}(\hat{\beta}))$$

Hypothesis Testing for Beta

- In the logistic regression model, the standard error of the maximum likelihood estimate $\hat{\beta}$ can also be estimated using calculus and likelihood-based methods

- The standard error of $\hat{\beta}$ is necessary for hypothesis testing and confidence intervals

Hypothesis Testing for Beta

- A 95% confidence interval for exp(β), the population odds ratio, is:

$$(\exp[\hat{\beta} - 1.96 \text{ s.e.}(\hat{\beta})], \exp[\hat{\beta} + 1.96 \text{ s.e.}(\hat{\beta})])$$

- If this interval contains the value 1, then the relationship between disease and exposure is not statistically significant

Hypothesis Testing using software!

Using any standard package, one would find:

- $\hat{\alpha} = -1.126687, s.e.(\hat{\alpha}) = 0.371386$
- $z = -30.34, P < 0.001$
- $95\%\ CI = (-1.199477, -1.053896)$
- $\hat{\beta} = -0.2832076, s.e.(\hat{\beta}) = 0.0713668$
- $z = -3.97, P < 0.001$
- $95\%\ CI = (-0.4230838, -0.1433313)$

Fitting Logistic Regression Models

- We start with

$$\textbf{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

- To make inferences and predictions, we need to estimate the population coefficients $\beta_0, \beta_1, \ldots, \beta_p$ using the information contained in a sample of observations.

- We don't use least squares for this, we use the *method of maximum likelihood*.

- The *likelihood* of a set of parameter values $(\beta_0, \beta_1, \ldots, \beta_p)$ given an observed sample of outcomes (data) is equal to the probability of those outcomes given the parameter values.

Maximum Likelihood

- The likelihood (or likelihood function) is defined by:

$$\boldsymbol{L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} L_i\,(\beta_0, \beta_1, \ldots, \beta_p)}$$

- Here we have a product of terms. *L* can be interpreted as the overall probability of obtaining the sample outcomes we observed given the regression parameters β_0, β_1, …, and β_p

- L_i is the contribution to the likelihood for subject *i*

- We want *L* to be large ("maximized")

Maximum Likelihood

- In the likelihood function, L_i is the contribution to the likelihood for subject *i*

$L_i = p_i$ for a subject with disease ($Y_i = 1$)

$L_i = 1 - p_i$ for a subject without disease ($Y_i = 0$)

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

$$1 - p_i = 1 / [1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})]$$

- The method of maximum likelihood will find estimated values of β_0, β_1, …, β_p which maximize the likelihood function $L(\beta_0, \beta_1, \ldots, \beta_p)$

Maximum Likelihood

$$L(\beta_0, \beta_1, \ldots, \beta_p)$$

$$= \prod_{i=1}^{n} L_i\left(\beta_0, \beta_1, \ldots, \beta_p\right)$$

$$= \prod_{i=1}^{n} p_i{}^{y_i}(1-p_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}\right)^{y_i} \cdot \left(\frac{1}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})^{y_i}}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}$$

where $y_i = 1$ if a subject has the disease and $y_i = 0$ if a subject does not have the disease

Maximum Likelihood

- We tend to take the natural log of the likelihood function which is easier to work with
- In general, we cannot maximize the likelihood function with a closed form solution, and iterative techniques are used to find the MLEs using a computer package
- The maximized likelihood is also used in the *likelihood ratio test*, comparing one simpler model nested within a more complex model

Maximum Likelihood

- We take derivatives of the log(L) function with respect to β_0 and β_1, set the derivatives equal to 0, and solve for β_0 and β_1
- We will call these estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Additional calculus and matrix methods can be used to estimate the *standard error* of the maximum likelihood estimate of each parameter and the *variance/covariance matrix* of the whole parameter vector
- These are necessary for hypothesis testing and confidence intervals

Maximum Likelihood

- The coefficient β_1 is the natural log of the odds ratio (odds of disease among exposed versus odds of disease among unexposed), which has an approximate normal distribution

- An approximate 95% (Wald) confidence interval for β_1 is:

$$(\hat{\beta}_1 - 1.96\ \text{se}(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96\ \text{se}(\hat{\beta}_1))$$

Ahead:

- Model building in logistic regression
- Regression diagnostics
- Goodness of fit and model validation

…and more examples!