Berkeley School of Public Health

Univariable and bivariable analyses of epidemiologic data

PHW250B

Epidemiologic analysis topics (this week)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses for
    - Linear regression
    - Logistic regression
    - Log-linear regression
- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data
  - Matched data

This video

Epidemiologic analysis topics (next week)

- Modern analytic approaches from the causal inference literature
  - Propensity score matching
  - Inverse probability of treatment weighting
  - G-computation
  - Double robust estimation
  - Machine learning

Big picture

- Considerations when planning statistical analysis of an epidemiologic study:
  - What is the research question?
  - What is the study design?
  - What type of data was collected?
    - Binary, continuous, categorical
  - What type of measure of disease was estimated?
    - Prevalent cases, incident cases
  - What are the threats to validity?
  - What is the desired measure of association?
- It is best for these considerations to drive the choice of the statistical analysis rather than vice versa.

Purpose of these videos

- Provide you with an overview of common statistical analysis approaches used by epidemiologists
  - Focus on intuition, not on statistical details
- The level of detail in these videos will help you become improve your skills in critically reviewing and interpreting published studies.
- To learn how to implement these methods, take additional statistics courses.

Types of data

Categorical data	Examples
Nominal data: data with levels or strata that do not have a meaningful order	Race categories *(African American, Asian American, Latino, Native American, White, Other)*
Binary data: nominal data with only two levels	Sex *(Male, Female)*
Ordinal data: data with levels or strata that do have a meaningful order	Health rating *(Excellent, good, fair, poor, very poor)*
Continuous data: data with ordered values and an equal distance between each level	Height in centimeters Age in years

Types of analyses

- **Univariable:** analysis of one variable (exposure or outcome)
  - Example: estimate the prevalence of disease
- **Bivariable:** analysis of two variables (exposure and outcome)
  - Example: estimate the crude relative risk for exposure and disease
- **Multivariable:** analysis of more than two variables (exposure, outcome, and confounders or other variables)
  - Example: estimate the relative risk for exposure and disease adjusting for sex, age, and race

Types of analyses

- Often we will conduct all three and report them in sequence in a publication.
- Univariable analyses provide important summaries of the magnitude of disease in a population
- Common example of comparing bivariable and multivariable analyses:
  - Compare crude RR to adjusted RR to assess the presence of confounding

Dependent and independent variables

- **Dependent variable**
  - This is the general term for the variable that we are focused on studying. We expect that the dependent variable will change based on the values of the independent variable.
  - In this video, we will refer to it as the **outcome**.
- **Independent variable**
  - Ideally this is a variable that is unaffected by any other variables and that may affect the dependent variable.
    - This is true in a trial but not necessarily in an observational study, but it is still called an independent variable in both types of studies.
  - In this video, we will refer to it as the **exposure**.

Case study in this video: WASH Benefits

- We will examine examples of univariable and bivariable analyses in this trial.
- The research question drove the study design and choice of outcome measures, which drove the choice of statistical analyses.
- The publication reported univariable, bivariable, and multivariable analyses.

Articles

Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial

Stephen P Luby, Mahbubur Rahman, Benjamin F Arnold, Leanne Unicomb, Sania Ashraf, Peter J Winch, Christine P Stewart, Farzana Begum, Faruqe Hussain, Jade Benjamin-Chung, Elli Leontsini, Abu M Naser, Sarker M Parvez, Alan E Hubbard, Audrie Lin, Fosiul A Nizame, Kaniz Jannat, Ayse Ercumen, Pavani K Ram, Kishor K Das, Jaynal Abedin, Thomas F Clasen, Kathryn G Dewey, Lia C Fernald, Clair Null, Tahmeed Ahmed, John M Colford Jr

Summary

Background Diarrhoea and growth faltering in early childhood are associated with subsequent adverse outcomes. We aimed to assess whether water quality, sanitation, and handwashing interventions alone or combined with nutrition interventions reduced diarrhoea or growth faltering.

Methods The WASH Benefits Bangladesh cluster-randomised trial enrolled pregnant women from villages in rural Bangladesh and evaluated outcomes at 1-year and 2-years' follow-up. Pregnant women in geographically adjacent clusters were block-randomised to one of seven clusters: chlorinated drinking water (water); upgraded sanitation (sanitation); promotion of handwashing with soap (handwashing); combined water, sanitation, and handwashing; counselling on appropriate child nutrition plus lipid-based nutrient supplements (nutrition); combined water, sanitation, handwashing, and nutrition; and control (data collection only). Primary outcomes were caregiver-reported diarrhoea in the past 7 days among children who were in utero or younger than 3 years at enrolment and length-for-age *Z* score among children born to enrolled pregnant women. Masking was not possible for data collection, but analyses were masked. Analysis was by intention to treat. This trial is registered at ClinicalTrials.gov, number NCC01590095.

Findings Between May 31, 2012, and July 7, 2013, 5551 pregnant women in 720 clusters were randomly allocated to one of seven groups. 1382 women were assigned to the control group; 698 to water; 696 to sanitation; 688 to handwashing; 702 to water, sanitation, and handwashing; 699 to nutrition; and 686 to water, sanitation, handwashing, and nutrition. 331 (6%) women were lost to follow-up. Data on diarrhoea at year 1 or year 2 (combined) were available for 14 425 children (7331 in year 1, 7094 in year 2) and data on length-for-age *Z* score in year 2 were available for 4584 children (92% of living children were measured at year 2). All interventions had high adherence. Compared with a prevalence of 5·7% (200 of 3517 child weeks) in the control group, 7-day diarrhoea prevalence was lower among index children and children under 3 years at enrolment who received sanitation (61 [3·5%] of 1760; prevalence ratio 0·61, 95% CI 0·46–0·81), handwashing (62 [3·5%] of 1795; 0·60, 0·45–0·80), combined water, sanitation, and handwashing (74 [3·9%] of 1902; 0·69, 0·53–0·90), nutrition (62 [3·5%] of 1766; 0·64, 0·49–0·85), and combined water, sanitation, handwashing, and nutrition (66 [3·5%] of 1861; 0·62, 0·47–0·81); diarrhoea prevalence was not significantly lower in children receiving water treatment (90 [4·9%] of 1824; 0·89, 0·70–1·13). Compared with control (mean length-for-age *Z* score −1·79), children were taller by year 2 in the nutrition group (mean difference 0·25 [95% CI 0·15–0·36]) and in the combined water, sanitation, handwashing, and nutrition group (0·13 [0·02–0·24]). The individual water, sanitation, and handwashing groups, and combined water, sanitation, and handwashing group had no effect on linear growth.

Lancet Glob Health 2018; 6: e302–15

Published Online January 29, 2018 http://dx.doi.org/10.1016/S2214-109X(17)30490-4

See Comment page e236

See Articles page e316

Division of Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA, USA (Prof S P Luby MD); **International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh** (M Rahman MBBS, L Unicomb PhD, S Ashraf MPH, F Begum MPH, F Hussain MSS, A M Naser MBBS, S M Parvez MPH, F A Nizame MA, K Jannat MBBS, K K Das MS, J Abedin MS, T Ahmed PhD); **School of Public Health University of California Berkeley, Berkeley, CA, USA** (B F Arnold PhD, J Benjamin-Chung PhD, Prof A L Hubbard PhD, A Lin PhD, A Ercumen PhD, Prof L C Fernald, Prof J M Colford Jr MD); **Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA** (Prof P J Winch MD, E Leontsini MD); **Department of Nutrition, University of**

Univariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	0	N/A	Univariable	Mean, median
Categorical	0	N/A	Univariable	Absolute number, percentage, prevalence, incidence

- **Example for continuous outcome:** height-for-age Z-score
  - Mean height-for-age Z-score at specific ages
- **Example for categorical outcome:** diarrhea status (yes / no)
  - Prevalence of diarrhea at a specific age

Example of univariable analysis in WASH Benefits

	N	Mean* prevalence
Control *vs* intervention		
Control	3517	5·7%
Water	1824	4·9%
Sanitation	1760	3·5%
Handwashing	1795	3·5%
Water, sanitation, and handwashing	1902	3·9%
Nutrition	1766	3·5%
Water, sanitation, handwashing, and nutrition	1861	3·5%

Table 4: **Diarrhoea prevalence 1 and 2 years (combined) after intervention**

Bivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Continuous	Bivariable	Correlation test
Categorical	1	Categorical	Bivariable	Chi-square test
Continuous	1	Binary	Bivariable	T-test
Continuous	1	Categorical	Bivariable	Analysis of Variance (ANOVA)
Categorical	1	Continuous or categorical	Bivariable	Simple logistic or log-linear regression

Bivariable analyses

Continuous exposure and outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Continuous	Bivariable	Correlation test

- Many types of correlation tests.
- **Question asked by common correlation tests:**
  - How does the relationship between the exposure and outcome compare relative to a straight line
  - Example: How does the mean weight-for-age and mean height-for-age compare at a specific age?

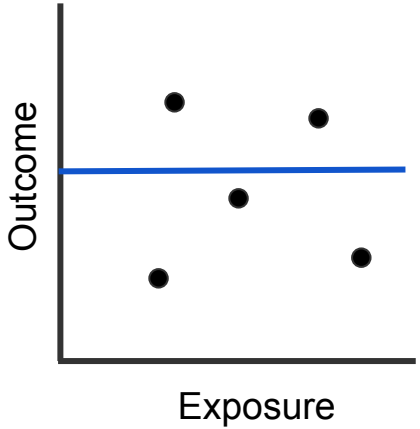
Bivariable analyses

Continuous exposure and outcome

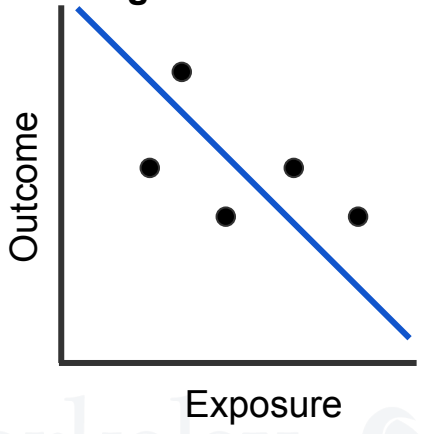
Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Continuous	Bivariable	Correlation test

Positive correlation

No correlation



Negative correlation



Bivariable analyses

Categorical exposure and outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Categorical	1	Categorical	Bivariable	Chi-square test

- **Question asked by Chi-square test for independence:**
  - Is the proportion of the outcome the same across levels of the exposure?
  - Example: Is the prevalence of diarrhea the same across different intervention arms in the WASH Benefits trial?
- Note: the chi-square test for independence is different from the chi-square test for homogeneity (used to assess the presence of effect modification)

Bivariable analyses

Binary exposure, continuous outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Binary	Bivariable	T-test

- **Question asked by t-test:**
  - Is the mean of the outcome the same across levels of the exposure?
  - Example: Is the mean height-for-age Z-score the same across different intervention arms in the WASH Benefits trial?
    - Considered binary because we are usually comparing an intervention arm to a control arm.

Bivariable analyses

Categorical exposure, continuous outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	1	Categorical	Bivariable	Analysis of Variance (ANOVA)

- **Question asked by ANOVA:**
  - Is the mean of the outcome the same across levels of the exposure?
  - Example: Is the mean height-for-age Z-score the same across different race categories?

Bivariable analyses

Continuous exposure, categorical outcome

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Categorical	1	Continuous or categorical	Bivariable	Simple logistic or log-linear regression

- **Question asked by logistic or log-linear regression:**
  - **Logistic**: Is the log odds of the outcome the same for every unit change in the exposure variable?
  - **Log-linear**: Is the log of the outcome the same for every unit change in the exposure variable?
  - Example: Is stunting prevalence the same across different race categories?

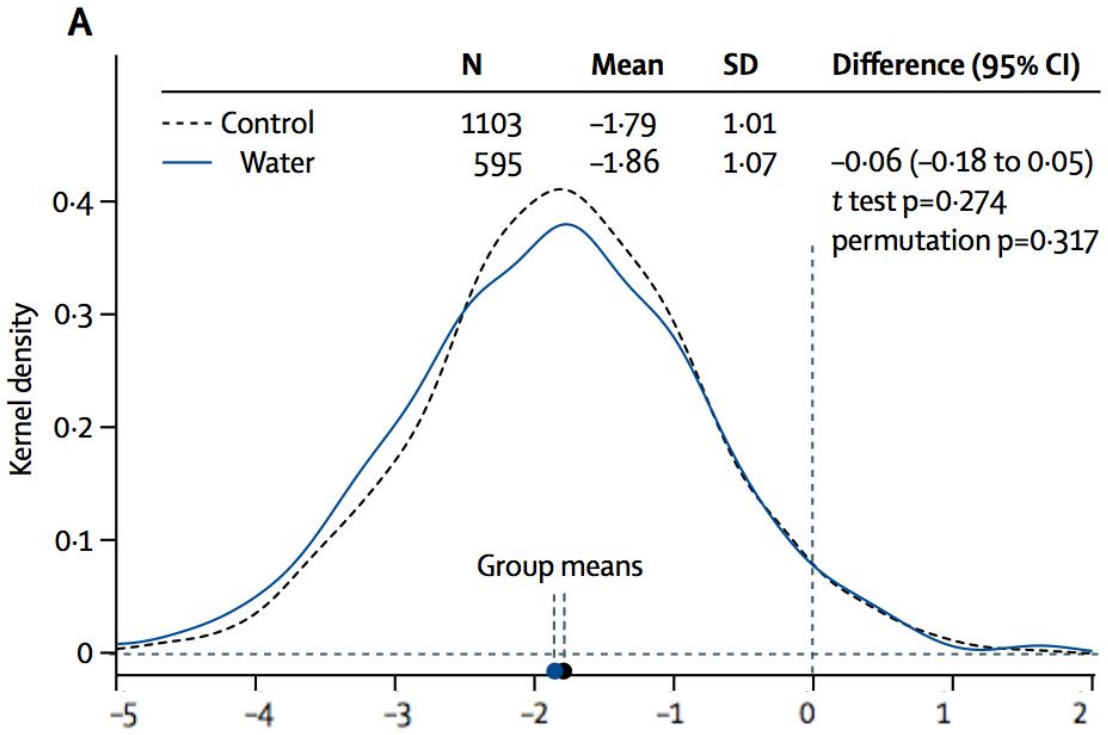
Example of bivariable analyses in WASH Benefits

	N	Mean* prevalence	Unadjusted† prevalence difference (95% CI)
Control *vs* intervention			
Control	3517	5·7%	..
Water	1824	4·9%	−0·6 (−1·9 to 0·6)
Sanitation	1760	3·5%	−2·2 (−3·4 to −1·0)
Handwashing	1795	3·5%	−2·3 (−3·4 to −1·1)
Water, sanitation, and handwashing	1902	3·9%	−1·7 (−2·9 to −0·6)
Nutrition	1766	3·5%	−2·0 (−3·1 to −0·8)
Water, sanitation, handwashing, and nutrition	1861	3·5%	−2·2 (−3·3 to −1·0)

Table 4: **Diarrhoea prevalence 1 and 2 years (combined) after intervention**

Example of bivariable analyses in WASH Benefits

Figure 3: **Intervention effects on length-for-age *Z* scores in 4584 children after 2 years of intervention**



Summary of univariable and bivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method
Continuous	0	N/A	Univariable	Mean, median
Categorical	0	N/A	Univariable	Absolute number, percentage
Continuous	1	Continuous	Bivariable	Correlation test
Categorical	1	Categorical	Bivariable	Chi-square test
Continuous	1	Binary	Bivariable	T-test
Continuous	1	Categorical	Bivariable	Analysis of Variance (ANOVA)
Categorical	1	Continuous or Categorical	Bivariable	Simple logistic or log-linear regression

Summary of key points

- It is best for the study design, type of data, threats to validity, and desired measure of association to drive the choice of the statistical analysis rather than vice versa.
- Often studies will conduct univariable, bivariable, and multivariable analyses and report them in sequence in a publication.
- Most commonly, univariable analyses provide information about the measure of disease or exposure on its own and bivariable analyses assess crude relationships between exposures and outcomes.

Berkeley School of Public Health

Multivariable linear regression models

PHW250B

Epidemiologic analysis topics (this week)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses
    - Linear regression
    - Logistic regression
    - Log-linear regression

  } This video

- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data
  - Matched data

} This video (brief introduction)

What we hope you learn on this topic in this course

- Identify which type of model is used depending on the type of dependent and independent variable
- How to interpret the coefficients from commonly used regression models
- Which type of regression models can be used to obtain each type of measure of association (mean difference, risk difference, risk ratio, rate ratio, odds ratio)
- Articulate certain assumptions of these models

Regression models in this video

- Are appropriate when the data are independent
- In other words: **No clustering or auto-correlation**
- **Examples**
  - **Example of clustered data:** Influenza is likely to be clustered within households since household members are likely to transmit it to each other
  - **Example of auto-correlation:** Longitudinal data with body mass index (BMI) measured on the same individual multiple times in a year
  - Different statistical approaches than the ones shown in this video must be used in these cases.

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio

Note: multivariable regression is also referred to as multiple regression

Linear regression

$$E(Y|X = x) = \underbrace{\beta_0 + \beta_1 x}$$

Outcome or dependent variable

Exposure or independent variable

Coefficients

Note: as written, this is technically a bivariable model. The next slide shows a multivariable model.

- For a binary exposure X, β_1 = the difference in the mean of the outcome Y when $X = 1$ and when $X = 0$
- Uses the **identity link function**: we model the outcome directly, without any transforming function
- Most common for continuous outcomes
  - Can use to model risks, rates and counts but combinations of betas produce impossible risk, rate and count values (e.g., negative values, values > 1 for risk)
- This model assumes no interaction.

How the exposure is coded

- **Categorical exposures**
  - Nominal: Create indicator variables for each category
  - Ordinal: Create a numbered categorical variable
- **Continuous exposures**
  - Often no need to recode
  - Can also categorize using the approaches above

How indicator variables are created

- **Categorical exposure example:**
  - City of residence

	X1	X2	X3	X4	X5
Original value					
Berkeley (reference)	0	0	0	0	0
Oakland	1	0	0	0	0
Albany	0	1	0	0	0
Piedmont	0	0	1	0	0
Emeryville	0	0	0	1	0
Kensington	0	0	0	0	1

How indicator variables are created

Ordinal exposure example:

- Highest education level measured with three values:

Original value	Value in variable used in regression model
Less than high school	1
Completed high school	2
More than high school	3

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

2. We use the notation from the model to write the formula for the mean difference.

$$E(Y|X = 1) - E(Y|X = 0)$$

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

2. We use the notation from the model to write the formula for the mean difference.

$$E(Y|X = 1) - E(Y|X = 0)$$

3. Next we fill in the coefficients for each term in the difference.

$$E(Y|X = 1) - E(Y|X = 0) = \big(\beta_0 + \beta_1 \cdot (1)\big) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

Obtaining the mean difference from linear regression coefficients

1. Start with the linear model specification

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

2. We use the notation from the model to write the formula for the mean difference.

$$E(Y|X = 1) - E(Y|X = 0)$$

3. Next we fill in the coefficients for each term in the difference.

$$E(Y|X = 1) - E(Y|X = 0) = \big(\beta_0 + \beta_1 \cdot (1)\big) - \big(\beta_0 + \beta_1 \cdot (0)\big)$$

4. After simplifying, we see that β_1 is equal to the mean difference.

$$\beta_1 = E(Y|X = 1) - E(Y|X = 0)$$

Example of multivariable linear regression

- **Outcome**: “height-for-age” (continuous)
  - Z-score with values ranging from -3 to 3
- **Exposure**: “sanitation” (binary)
  - Indicator variable with the values:
    - 0 = unimproved sanitation
    - 1 = improved sanitation
- **Confounder**: “wealth” (binary)
  - Indicator variable with the values:
    - 0 = below median household wealth
    - 1 = above median household wealth

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- The independent variables in this model are Sanitation and Wealth.
- These are often referred to as **covariates**, a term which includes the exposure / intervention variable and potential confounders and other variables that are adjusted for.

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- β_0 = Mean height-for-age z-score among those without improved sanitation and with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) = \beta_0$$

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1\text{Sanitation} + \beta_2\text{Wealth}$$

- β_0 = Mean height-for-age z-score among those without improved sanitation and with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) = \beta_0$$

- β_1 = Difference in mean height-for-age z-score among people with vs. without improved sanitation with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 1, \text{Wealth} = 0) - E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) =$$
$$\big(\beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0)\big) - \big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0)\big) = \beta_1$$

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

- β_0 = Mean height-for-age z-score among those without improved sanitation and with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) = \beta_0$$

- β_1 = Difference in mean height-for-age z-score among people with vs. without improved sanitation with below median household income

$$E(\text{Height-for-age}|\text{Sanitation} = 1, \text{Wealth} = 0) - E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) =$$
$$\big(\beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0)\big) - \big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0)\big) = \beta_1$$

- β_2 = Difference in mean height-for-age z-score among people in below vs. above median household income categories among those without improved sanitation

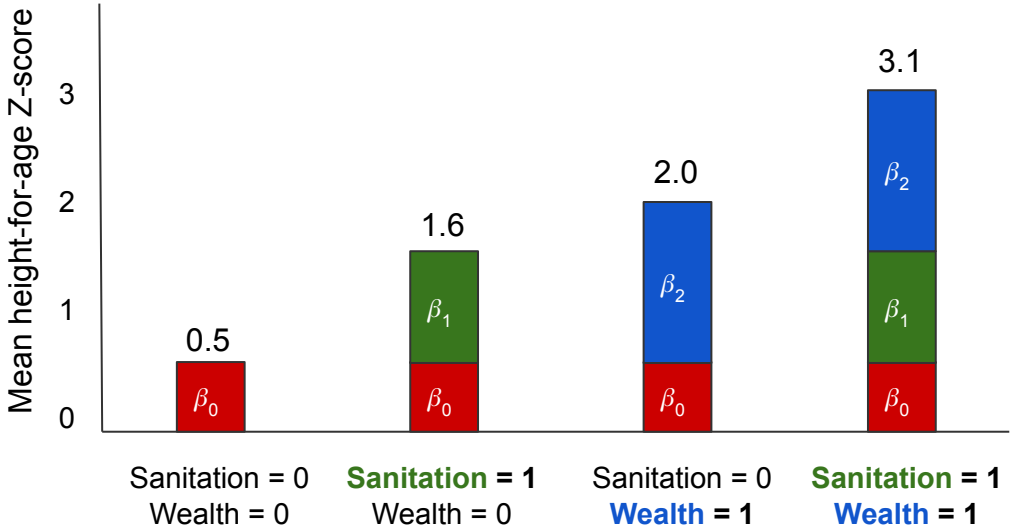
$$E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 1) - E(\text{Height-for-age}|\text{Sanitation} = 0, \text{Wealth} = 0) =$$
$$\big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1)\big) - \big(\beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0)\big) = \beta_2$$

Example of multivariable linear regression

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = \beta_0 + \beta_1 \text{Sanitation} + \beta_2 \text{Wealth}$$

$$E(\text{Height-for-age}|\text{Sanitation, Wealth}) = 0.5 + 1.1 \cdot \text{Sanitation} + 1.5 \cdot \text{Household wealth}$$

- Mean difference for **sanitation** alone: 1.6 - 0.5 = 1.1
- Mean difference for above median **wealth** alone: 2.0 - 0.5 = 1.5
- Mean difference for **sanitation** + above median **wealth**: (1.5 + 1.1 + 0.5) - 0.5 = 2.6



Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- Continuous independent variables are interpreted as the association of the dependent variable (the outcome) with a one-unit change in the independent variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one-kilogram change in weight from 0 kg to 1 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (1)) - (-2.5 + 1.1 (0) + 0.05 (0)) = 0.05

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- Continuous independent variables are interpreted as the association of the dependent variable (the outcome) with a one-unit change in the independent variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one-kilogram change in weight from 0 kg to 1 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (1)) - (-2.5 + 1.1 (0) + 0.05 (0)) = 0.05
- … **for a one-kilogram change in weight from 10 kg to 11 kg?**
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (11)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 0.05

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- The coefficients for some continuous independent variables may not be easily interpretable, so it may be more appropriate to estimate, for example, a 1 SD increase in the continuous variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one SD (20 kg) change in weight?**

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- The coefficients for some continuous independent variables may not be easily interpretable, so it may be more appropriate to estimate, for example, a 1 SD increase in the continuous variable.
- **Example:** What is the difference in mean height-for-age (HAZ) Z-score among those without improved sanitation (Sanitation = 0) **for a one SD (20 kg) change in weight?**
  - If we are holding everything else constant (e.g., Sanitation) then we can just multiply the coefficient on weight times the SD.
  - Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (30)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 1
  - More simply: 0.05 * 20 = 1

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

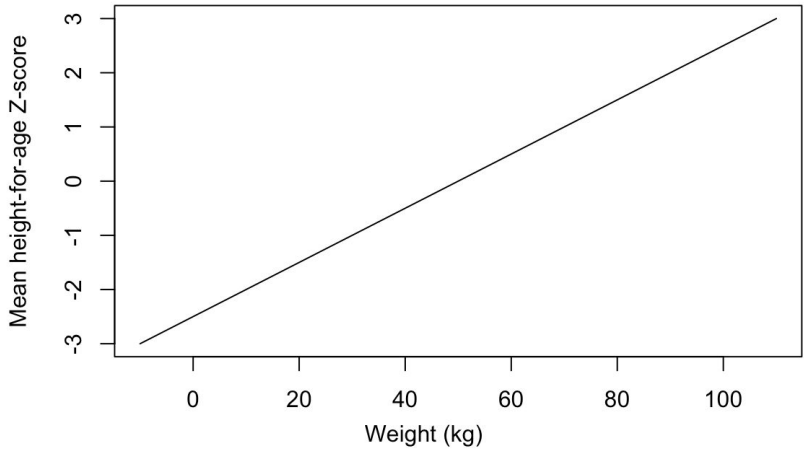
- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables

Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables

- Plot of the mean height-for-age Z-score when Sanitation = 0
- y = mx + b
  - y = dependent variable
  - m = slope = 0.05
  - x = independent variable
  - b = intercept = -2.5

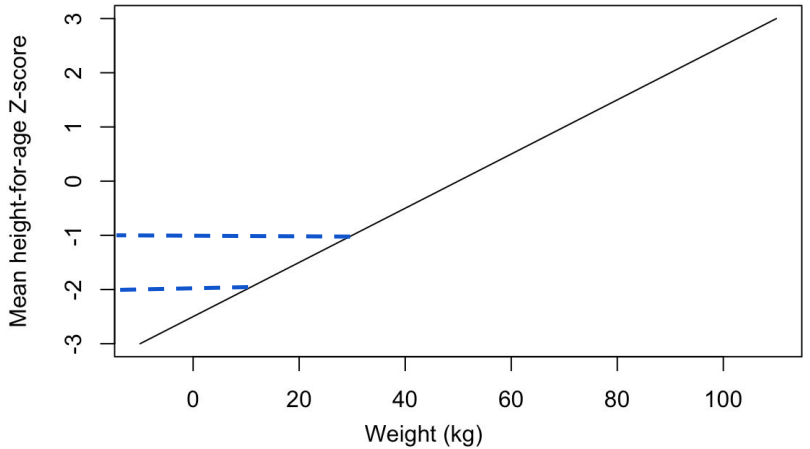


Interpreting linear regression coefficients for a continuous variable

$$E(\text{Height-for-age}|\text{Sanitation, Weight}) = -2.5 + 1.1 \cdot \text{Sanitation} + 0.05 \cdot \text{Weight}$$

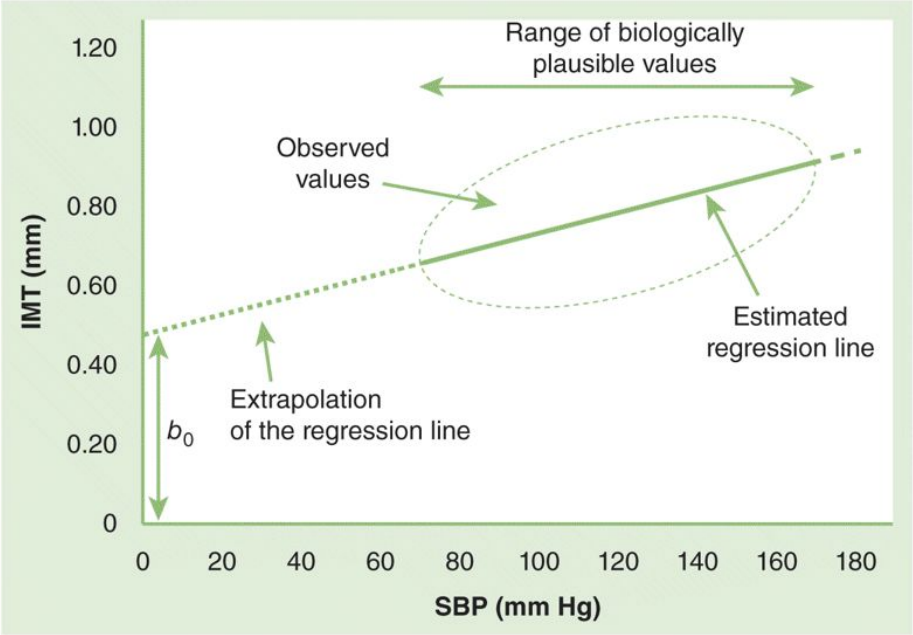
- This illustrates one of the assumptions of linear regression models:
  - There is a **linear relationship** between independent and dependent variables

- Plot of the mean height-for-age Z-score when Sanitation = 0
- y = mx + b
  - y = dependent variable
  - m = slope = 0.05
  - x = independent variable
  - b = intercept = -2.5



- From last slide: Mean difference in HAZ = (-2.5 + 1.1 (0) + 0.05 (30)) - (-2.5 + 1.1 (0) + 0.05 (10)) = 1
  - (-2.5 + 1.1 (0) + 0.05 (30)) = -1
  - (-2.5 + 1.1 (0) + 0.05 (10)) = -2

Extrapolation in regression models



- Linear regression models assume a linear relationship between X and Y.
- The intercept is an extrapolation of the regression line to the y-axis.
- In some cases, this is beyond the plausible range of values.
- It is best to be very cautious when interpreting regression results extrapolated beyond the range of the observed data.

Strengths / weaknesses of bivariable vs multivariable analyses

- **Bivariable analyses**
  - Simple to perform
  - Cannot adjust for confounding
- **Multivariable analyses**
  - Can adjust for confounding
  - More complex to perform
  - When there are many covariates, strata may become very sparse, requiring extrapolation beyond the data using the model.
  - Typically require making assumptions about our data that may not be possible to validate

Regression models for repeated measures data

- Models so far have assumed data has **no clustering or auto-correlation**
- Models also exist that **allow for clustering or auto-correlation**
- Examples of studies with this type of data:
  - Cluster-randomized studies
    - Outcomes and exposures of people in the same cluster may be statistically dependent.
    - Additional data from people in the same cluster adds less information than data from people in different clusters.
  - Cohort studies with repeated measurements collected longitudinally
    - Measurements on the same person over time are likely to be auto-correlated — e.g., a person's weight on Monday is correlated with their weight on Tuesday

Data in a repeated measures study

Data structure for single measurement

id	weight
1	60
2	80
3	55
4	62
5	81

Data structure for repeated measurement

id	time	weight
1	1	60
1	2	62
2	1	80
2	2	76
3	1	55
3	2	54

Regression models for repeated measures data

- As is true for data with single measurements per individual, with repeated measures data, the type of outcome and desired measure of association determines whether we use a linear, logistic, or log-linear model.
- Whenever there are repeated measures on individuals, the data has to be analyzed accounting for correlation among observations within individuals
- More advanced courses cover the specific models appropriate for these types of data:
  - Generalized linear models with robust standard errors
  - Mixed models (or “random effects” models)
  - Generalized estimating equations

Summary of key points

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Continuous outcome**: linear regression
- **How to interpret the coefficients from this model**
  - $E(Y|X = x) = \beta_0 + \beta_1 x$
  - Mean difference = β_1
- **Key assumptions of linear models:** no clustering or autocorrelation

Berkeley School of Public Health

Interaction in multivariable analyses of epidemiologic data

PHW250B

Epidemiologic analysis topics (this week)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses
    - Linear regression
    - Logistic regression
    - Log-linear regression

    } This video
- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data
  - Matched data

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- So far the models we have looked at have assumed there is no interaction.
- How do we specify potential interaction in our model if we suspect that it exists and want to investigate it?
- To assess possible interaction using a regression model, we include a term in the model that is the product of two other covariates.
- In the model above, β_3 assess the interaction between x_1 and x_2.

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

	x_1	x_2	$x_1 * x_2$
x_1 present, x_2 absent	1	0	0
x_1 absent, x_2 present	0	1	0
x_1 present, x_2 present	1	1	1
x_1 absent, x_2 absent	0	0	0

- β_0 = mean of Y when x_1 and x_2 are absent
- $\beta_0 + \beta_1$ = mean of Y when x_1 is present and x_2 is absent
- $\beta_0 + \beta_2$ =mean of Y when x_1 is absent and x_2 is present
- $\beta_0 + \beta_1 + \beta_2 + \beta_3$ = mean of Y when x_1 and x_2 are both present

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

	x_1	x_2	$x_1 * x_2$
x_1 present, x_2 absent	1	0	0
x_1 absent, x_2 present	0	1	0
x_1 present, x_2 present	1	1	1
x_1 absent, x_2 absent	0	0	0

- β_1 = Mean difference *Y* between those with $x_1 = 0$ and $x_1 = 1$ among those with $x_2 = 0$
- β_2 = Mean difference *Y* between those with $x_2 = 0$ and $x_2 = 1$ among those with $x_1 = 0$
- β_3 = Mean additional difference in the value of *Y* beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference *Y* between those with x_1 = 0 and x_1 = 1 among those with x_2 = 0

1. Plug in values x_1 = 1 and x_2 = 0 into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference *Y* between those with x_1 = 0 and x_1 = 1 among those with x_2 = 0

1. Plug in values x_1 = 1 and x_2 = 0 into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

2. Plug in values x_1 = 0 and x_2 = 0 into the model

$$E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) + \beta_3 \cdot (0) \cdot (0)$$
$$E(Y|x_1 = 0, x_2 = 0) = \beta_0$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference *Y* between those with x_1 = 0 and x_1 = 1 among those with x_2 = 0

1. Plug in values x_1 = 1 and x_2 = 0 into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

2. Plug in values x_1 = 0 and x_2 = 0 into the model

$$E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) + \beta_3 \cdot (0) \cdot (0)$$
$$E(Y|x_1 = 0, x_2 = 0) = \beta_0$$

3. Take the difference in betas from Step 1 and 2

$$E(Y|x_1 = 1, x_2 = 0) - E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 - \beta_0$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_1 = Mean difference *Y* between those with x_1 = 0 and x_1 = 1 among those with x_2 = 0

1. Plug in values x_1 = 1 and x_2 = 0 into the model

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (0) + \beta_3 \cdot (1) \cdot (0)$$
$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$

2. Plug in values x_1 = 0 and x_2 = 0 into the model

$$E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (0) + \beta_3 \cdot (0) \cdot (0)$$
$$E(Y|x_1 = 0, x_2 = 0) = \beta_0$$

3. Take the difference in betas from Step 1 and 2

$$E(Y|x_1 = 1, x_2 = 0) - E(Y|x_1 = 0, x_2 = 0) = \beta_0 + \beta_1 - \beta_0$$

This equation is equivalent to the text above.

$$E(Y|x_1 = 1, x_2 = 0) - E(Y|x_1 = 0, x_2 = 0) = \beta_1$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_2 = Mean difference *Y* between those with $x_2 = 0$ and $x_2 = 1$ among those with $x_1 = 0$

Practice this on your own!

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of *Y* beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

1. From previous slide, the mean difference for x_1 when $x_2 = 0$

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of *Y* beyond the effect of x_1 among those with x_2 = 0 and the effect of x_2 among those with x_1 = 0

1. From previous slide, the mean difference for x_1 when x_2 = 0

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

2. Calculate the mean difference for x_2 when d x_1 = 1

$$E(Y|x_1 = 0, x_2 = 1) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1) + \beta_3 \cdot (0) \cdot (1)$$
$$\text{Mean Diff}_{01} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of *Y* beyond the effect of x_1 among those with x_2 = 0 and the effect of x_2 among those with x_1 = 0

1. From previous slide, the mean difference for x_1 when x_2 = 0

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

2. Calculate the mean difference for x_2 when d x_1 = 1

$$E(Y|x_1 = 0, x_2 = 1) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1) + \beta_3 \cdot (0) \cdot (1)$$
$$\text{Mean Diff}_{01} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

3. Calculate the mean difference for x_1 = 1 and x_2 = 1

$$E(Y|x_1 = 1, x_2 = 1) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (1) + \beta_3 \cdot (1) \cdot (1)$$
$$\text{Mean Diff}_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 = \beta_1 + \beta_2 + \beta_3$$

Regression models with interaction terms

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

- β_3 = Mean additional difference in the value of *Y* beyond the effect of x_1 among those with $x_2 = 0$ and the effect of x_2 among those with $x_1 = 0$

1. From previous slide, the mean difference for x_1 when $x_2 = 0$

$$E(Y|x_1 = 1, x_2 = 0) = \beta_0 + \beta_1 \cdot (1)$$
$$\text{Mean Diff}_{10} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

2. Calculate the mean difference for x_2 when d $x_1 = 1$

$$E(Y|x_1 = 0, x_2 = 1) = \beta_0 + \beta_1 \cdot (0) + \beta_2 \cdot (1) + \beta_3 \cdot (0) \cdot (1)$$
$$\text{Mean Diff}_{01} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

3. Calculate the mean difference for $x_1 = 1$ and $x_2 = 1$

$$E(Y|x_1 = 1, x_2 = 1) = \beta_0 + \beta_1 \cdot (1) + \beta_2 \cdot (1) + \beta_3 \cdot (1) \cdot (1)$$
$$\text{Mean Diff}_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 = \beta_1 + \beta_2 + \beta_3$$

Take the difference between equations 1+2 and equation 3.

$$\text{Mean Diff}_{11} - \text{Mean Diff}_{10} - \text{Mean Diff}_{01} =$$
$$= (\beta_1 + \beta_2 + \beta_3) - \beta_1 - \beta_2 = \beta_3$$

Scale of interaction and scale of model

- If we include an interaction term in an **additive** scale model (i.e., a model with an identity link), we assess interaction on the additive scale.
  - Linear regression
- If we include an interaction term in a **relative** scale model (i.e., a model with a log or logit link), we assess interaction on the relative / multiplicative scale.
  - Log-linear regression
  - Logistic regression
- To assess additive scale interaction from a relative scale model, we need to estimate the relative excess risk due to interaction (RERI).

Graphical demonstration of positive interaction

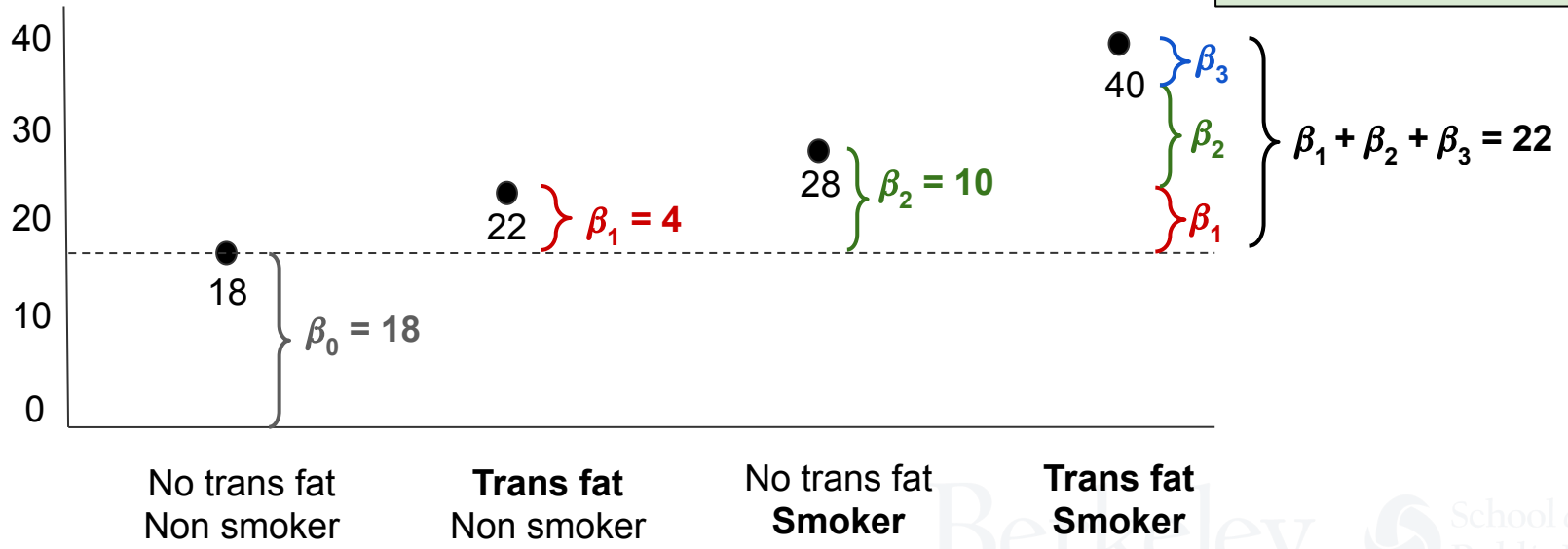
$$E(Y|x_1, x_2) = 18 + 4x_1 + 10x_2 + 8x_1 \cdot x_2$$

- Y: body mass index
- x_1: regularly consumes trans fat
- x_2: smokes cigarettes

Positive interaction because β_3 is positive.

β_3 = (22 - 10 - 4) = 8

$Diff_{11} > Diff_{10} + Diff_{01}$



Graphical demonstration of no interaction

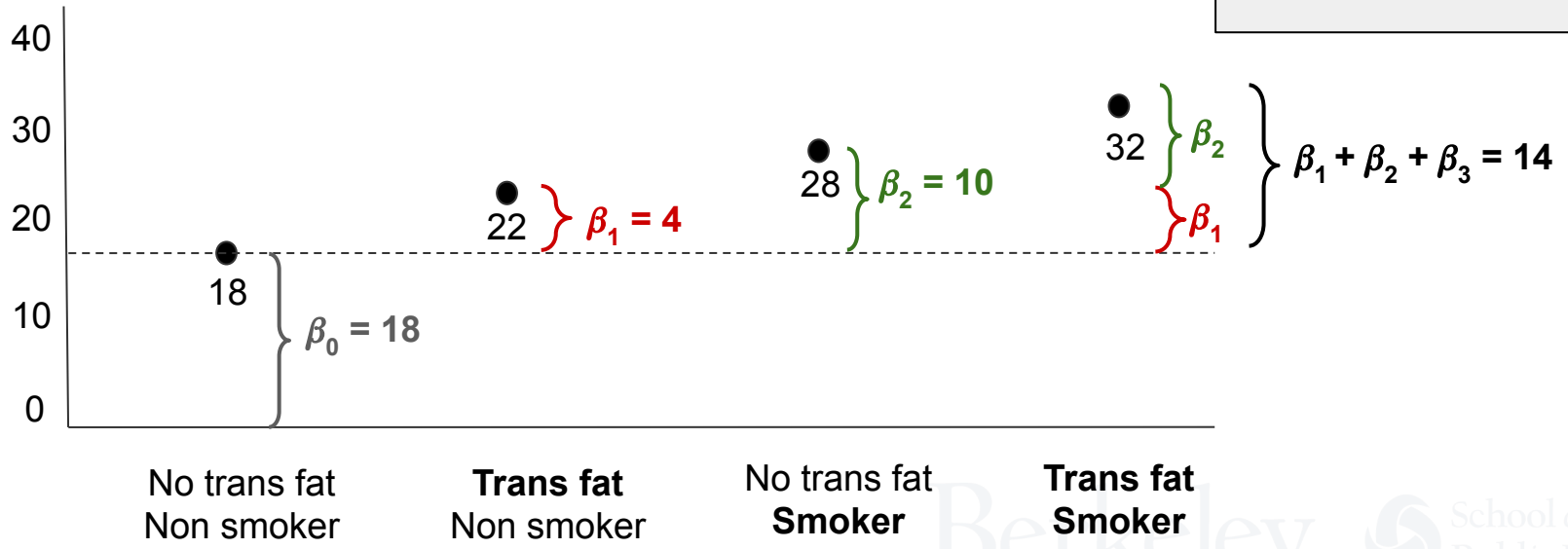
$$E(Y|x_1, x_2) = 18 + 4x_1 + 10x_2 + 0x_1 \cdot x_2$$

- Y: body mass index
- x_1: regularly consumes trans fat
- x_2: smokes cigarettes

No interaction because β_3 = 0.

β_3 = (14 - 10 - 4) = 0

$Diff_{11} = Diff_{10} + Diff_{01}$



Graphical demonstration of negative interaction

$$E(Y|x_1, x_2) = 18 + 4x_1 + 10x_2 - 7x_1 \cdot x_2$$

- Y: body mass index
- x_1: regularly consumes trans fat
- x_2: smokes cigarettes

Negative interaction because β_3 is negative.

β_3 = (7 - 10 - 4) = -7

$Diff_{11} < Diff_{10} + Diff_{01}$

P-values and interaction coefficients

- When we include interaction terms, we can obtain a p-value.
  - Standard statistical software packages return this by default.
- The p-value for the interaction term is analogous to p-value from the test for homogeneity when assessing interaction with stratification.

Summary of key points

- We can assess the interaction between two variables in our model by including the product of the two variables as a covariate in the model.
- The coefficient on the product estimates a quantity that compares the observed and expected joint association of two variables.
- The scale of the model is the scale that interaction is assessed on by default.
- The p-value for the interaction term is analogous to p-value from the test for homogeneity when assessing interaction with stratification.

Berkeley School of Public Health

Multivariable log-linear and logistic regression models

PHW250B

Epidemiologic analysis topics (this week)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses
    - Linear regression
    - Logistic regression
    - Log-linear regression

  } This video

- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data
  - Matched data

What we hope you learn on this topic in this course

- Identify which type of model is used depending on the type of dependent and independent variable
- How to interpret the coefficients from commonly used regression models
- Which type of regression models can be used to obtain each type of measure of association (mean difference, risk difference, risk ratio, rate ratio, odds ratio)
- Articulate certain assumptions of these models

Regression models in this video

- Are appropriate when the data are independent
- In other words: **No clustering or auto-correlation**
- Different statistical approaches than the ones shown in this video must be used in these cases.

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio

Note: multivariable regression is also referred to as multiple regression

Log-linear regression

$$\underbrace{\ln(E(Y|X=x))}_{\text{Log link}} = \underbrace{\beta_0 + \beta_1 x}_{\text{Coefficients}}$$

Note: you will see both **ln** and **log** used in the slides and the textbooks. In this course, you can always assume that **log** means **ln** (**log** with base e).

- For a binary exposure X, β_1 = the log risk or log rate ratio comparing the log risk or log rate when $X = 1$ and when $X = 0$
- Uses the **log link function**: we model the logarithm of the outcome as a function of the linear predictors
- Also known as “exponential risk” or “rate models”
- The log means the model can only predict positive values (no negative values).
- When modeling risk, some combinations of coefficients can lead to risks above 1, but typically not an issue for outcomes with relatively low risk.

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

3. Next we fill in the coefficients for each term in the difference in risk.

$$\ln(E(Y|X=1)) - \ln(E(Y|X=0)) = \beta_0 + \beta_1 \cdot (1) - \left(\beta_0 + \beta_1 \cdot (0)\right)$$

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

3. Next we fill in the coefficients for each term in the difference in risk.

$$\ln(E(Y|X=1)) - \ln(E(Y|X=0)) = \beta_0 + \beta_1 \cdot (1) - \left(\beta_0 + \beta_1 \cdot (0)\right)$$

4. After simplifying, we see that β_1 is equal to the natural log of the risk ratio.

$$\beta_1 = \ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right)$$

Obtaining the risk ratio (or rate ratio) from log-linear regression coefficients

1. Start with the log-linear model specification

$$\ln(E(Y|X=x)) = \beta_0 + \beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the risk as the difference in risk

$$\ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right) = \ln(E(Y|X=1)) - \ln(E(Y|X=0))$$

3. Next we fill in the coefficients for each term in the difference in risk.

$$\ln(E(Y|X=1)) - \ln(E(Y|X=0)) = \beta_0 + \beta_1 \cdot (1) - \left(\beta_0 + \beta_1 \cdot (0)\right)$$

4. After simplifying, we see that β_1 is equal to the natural log of the risk ratio.

$$\beta_1 = \ln\left(\frac{E(Y|X=1)}{E(Y|X=0)}\right)$$

5. When we exponentiate β_1, it equals the risk ratio.

$$\exp(\beta_1) = \frac{E(Y|X=1)}{E(Y|X=0)}$$

Types of log-linear regression models

- **Poisson model**
  - Uses a log link and assumes the outcome follows a Poisson distribution
- **Negative binomial model**
  - Uses a log link and assumes the outcome follows a negative binomial distribution
  - Makes less strong assumptions about the distribution of the variance than the Poisson model
- Both are commonly used with count data.
  - e^{β} is the relative association between mean counts when X=1 and X=0

Using log-linear regression models to estimate rate ratios

$$\ln\left(\frac{E(Y|X=x)}{\text{Person-time}}\right) = \beta_0 + \beta_1 x$$

$$\ln(E(Y|X=x)) - \ln(\text{Person-time}) = \beta_0 + \beta_1 x$$

$$\ln(E(Y|X=x)) = \ln(\text{Person-time}) + \beta_0 + \beta_1 x$$

- When the outcome is a count of events (e.g., a count of incident cases), we can estimate incidence rates by including a **person-time offset**.
- In log-linear models, this offset is the natural log of the person-time.

Example of log-linear regression

What risk factors are associated with incident coronary heart disease (CHD)?

TABLE 7-23 Data used for calculating the results shown in Table 7-22.

Cell	Male	Smok	Old age	Hyperten	Hypercho	Obese	PY	CHD	LogPY
1	0	0	0	0	0	0	1740.85	1	7.46213
2	0	0	0	0	0	1	1181.40	2	7.07446
3	0	0	0	0	1	0	539.97	0	6.29152
4	0	0	0	0	1	1	521.93	1	6.25754
…/…									
61	1	1	1	1	0	0	24.48	1	3.19804
62	1	1	1	1	0	1	37.41	0	3.62208
63	1	1	1	1	1	0	171.41	1	5.14405
64	1	1	1	1	1	1	85.37	5	4.44701

Szklo et al., 3rd Ed.

Example of log-linear regression

$$\ln(\text{incidence}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$\ln(\text{incidence}) = -6.3473 + 1.1852 x_1 + 0.6384 x_2 + 0.2947 x_3 + 0.5137 x_4 + 0.6795 x_5 + 0.2656 x_6$$

TABLE 7-22 Results from a Poisson regression analysis of binary predictors of coronary heart disease (CHD) incidence in the Washington County cohort of the Atherosclerosis Risk in Communities (ARIC) Study, ages 45–64 years at baseline, 1987–1994.

Variable (unit)	Poisson regression coefficient	Rate ratio
Intercept	−6.3473	—
Gender (male = 1, female = 0)	1.1852	3.27
Smoking (yes = 1, no = 0)	0.6384	1.89
Older age* (yes = 1, no = 0)	0.2947	1.34
Hypertension† (yes = 1, no = 0)	0.5137	1.67
Hypercholesterolemia‡ (yes = 1, no = 0)	0.6795	1.97
Obesity§ (yes = 1, no = 0)	0.2656	1.30

*Age ≥ 55 years.
†Blood pressure ≥ 140 mm Hg systolic or ≥ 90 mm Hg diastolic or antihypertensive therapy.
‡Total serum cholesterol ≥ 240 mg/dL or lipid-lowering treatment.
§Body mass index ≥ 27.8 kg/m^2 in males and ≥ 27.3 kg/m^2 in females.

Binary covariate

Incidence rate ratio for CHD for males compared to females is $e^{1.1852} = 3.27$

Szklo et al., 3rd Ed.

Logistic regression

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \underbrace{\text{logit}(E(Y|X=x))}_{\text{Logit link}} = \underbrace{\beta_0+\beta_1 x}_{\text{Coefficients}}$$

- For a binary exposure,
  - β_1 is the log odds ratio comparing the odds of the outcome when $X = 1$ and $X = 0$
  - $e^{\beta 1}$ is the odds ratio comparing the odds of the outcome when $X = 1$ and $X = 0$
- Uses the **logit link function**: we model the log-odds of the outcome as a function of the linear predictors
- The model only predicts values between 0 and 1.
  - This is the reason it is used so often even when measure of association under study is risk (not odds).
- Useful for case-control studies. The odds ratio produced by this model may estimate a CIR or IDR depending on the control sampling strategy.

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)$$

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)$$

3. Next we fill in the coefficients for each term in the difference in log odds.

$$\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)=\beta_0+\beta_1\cdot(1)-\big(\beta_0+\beta_1\cdot(0)\big)$$

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)$$

3. Next we fill in the coefficients for each term in the difference in log odds.

$$\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)=\beta_0+\beta_1\cdot(1)-\big(\beta_0+\beta_1\cdot(0)\big)$$

4. After simplifying, we see that β_1 is equal to the natural log of the odds ratio.

$$\beta_1=\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{OR})$$

Obtaining the odds ratio from logistic regression coefficients

1. Start with the logistic model specification

$$\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right)=\beta_0+\beta_1 x \qquad \ln(\text{Odds if } X=x)=\beta_0+\beta_1 x$$

2. Using log rules, we can write the natural log of the ratio of the odds as the difference in odds

$$\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)$$

3. Next we fill in the coefficients for each term in the difference in log odds.

$$\ln(\text{Odds if } X=1)-\ln(\text{Odds if } X=0)=\beta_0+\beta_1\cdot(1)-\left(\beta_0+\beta_1\cdot(0)\right)$$

4. After simplifying, we see that β_1 is equal to the natural log of the odds ratio.

$$\beta_1=\ln\left(\frac{\text{Odds if } X=1}{\text{Odds if } X=0}\right)=\ln(\text{OR})$$

5. When we exponentiate β_1, it equals the odds ratio.

$$\exp(\beta_1)=\text{OR}$$

Example of logistic regression

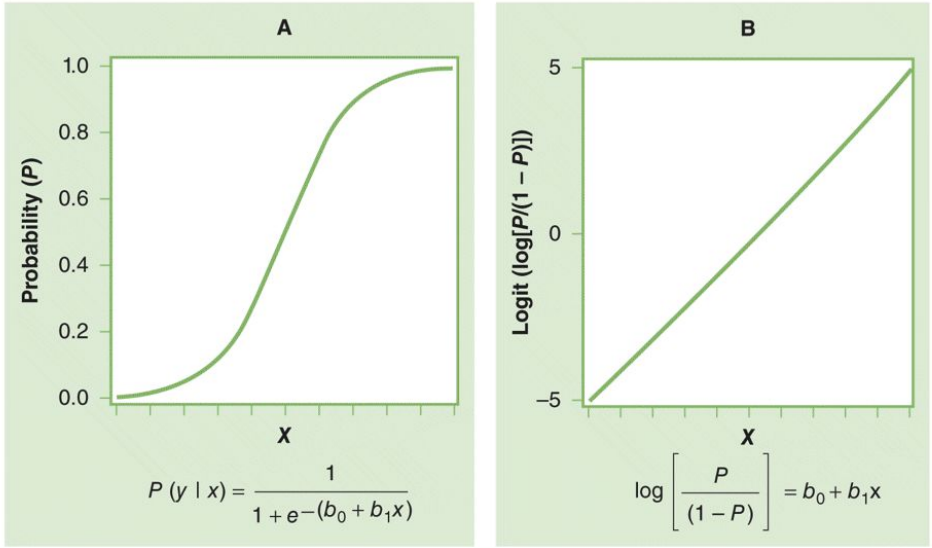


FIGURE 7-7 Two mathematically equivalent formulations of the logistic regression function.

Example of logistic regression

$$\ln(\text{odds ratio}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$\ln(\text{odds ratio}) = -8.9502 + 1.3075 x_1 + 0.7413 x_2 + 0.0114 x_3 + 0.0167 x_4 + 0.0074 x_5 + 0.0240 x_6$$

TABLE 7-18 Results from a logistic regression analysis of binary and continuous predictors of coronary heart disease (CHD) incidence in the Washington County cohort of the Atherosclerosis Risk in Communities (ARIC) Study, ages 45–64 years at baseline, 1987–1994.

Variable (unit)	Logistic regression coefficient	Odds ratio
Intercept	−8.9502	—
Gender (male = 1, female = 0)	1.3075	3.70
Smoking (yes = 1, no = 0)	0.7413	2.10
Age (1 year)	0.0114	1.011
Systolic blood pressure (1 mm Hg)	0.0167	1.017
Serum cholesterol (1 mg/dL)	0.0074	1.007
Body mass index (1 kg/m^2)	0.0240	1.024

Binary covariate

Odds ratio for CHD for males compared to females is $e^{1.3075} = 3.70$.

Continuous covariate

Odds ratio for CHD for a 1 mm Hg increase in systolic blood pressure is $e^{0.0167} = 1.017$.

Summary of key points

Log-linear models

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Count / rate / binary outcome:** log-linear regression
- **How to interpret the coefficients from commonly used regression models**
  - $\ln(E(Y|X = x)) = \beta_0 + \beta_1 x$
  - RR = $e^{\beta 1}$
  - The exponentiated β_1 ($e^{\beta 1}$) is the risk or rate ratio comparing the risk or rate when x = 1 to x = 0
- **Key assumptions of log-linear models:** no clustering or autocorrelation

Summary of key points

Logistic models

- Identify which type of model is used depending on the type of dependent and independent variable
  - **Binary outcome:** logistic regression
- **How to interpret the coefficients from commonly used regression models**
  - $\ln\left(\frac{E(Y|X=x)}{1-E(Y|X=x)}\right) = \beta_0 + \beta_1 x$
  - OR = $e^{\beta 1}$
  - The exponentiated β_1 ($e^{\beta 1}$) is the odds ratio comparing the odds when x = 1 to x = 0
- **Key assumptions of logistic models:** no clustering or autocorrelation

Berkeley School of Public Health

Analyses of survival data

PHW250B

Epidemiologic analysis topics (this week)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses
    - Linear regression
    - Logistic regression
    - Log-linear regression
- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data **This video**
  - Matched data

Summary of multivariable analyses

Type of outcome variable	Number of exposure and other variables	Type of exposure variable	Type of analysis	Analysis method	Measure of association estimated
Continuous	>1	Continuous or binary	Multivariable	Linear regression	Mean difference
Binary or count	>1	Continuous or binary	Multivariable	Log-linear regression	Risk ratio or rate ratio
Binary	>1	Continuous or binary	Multivariable	Logistic regression	Odds ratio
Time to event	>1	Continuous or binary	Multivariable	Cox proportional hazards model	Hazard ratio

Survival data

- Survival data includes measurements of the time that passed before a person developed a certain disease or condition.
- Recall the concept of the **hazard** — the instantaneous potential for change in disease status per unit of time at time t relative to the size of the candidate (i.e., disease-free) population at time *t*
- **Examples of survival or "time-to-event" data:**
  - Among pancreatic cancer patients, the time till death
  - Among drivers who use their mobile phone while driving, the time till a car accident
  - Time to recovery for patients receiving a new form of hip replacement

Hazard ratios

- Hazard ratios compare the hazard between the exposed and unexposed.
- To adjust hazard ratios for potential confounders, we can use **Cox proportional hazards models.**
- The hazard ratio from Cox proportional hazards regression approximates the RR and OR from log-linear or logistic regression when the cumulative risk is small.
- Then why use a survival approach?
  - Sometimes the question of interest is about the time to an event.
  - **Example**: How long does it take for patients to recover from influenza if they receive Tamiflu, an antiviral, in the first 3 days of onset?

Cox proportional hazards regression

The hazard is indexed by time (t) because it varies over time

$$\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$$

Baseline hazard

Coefficient

Cox proportional hazards regression

$$\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$$

Baseline hazard

Coefficient

$$h(t|X=x) = e^{(\ln h_0(t) + cx)} = h_0(t) \times e^{cx}$$

Cox proportional hazards regression

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

Baseline hazard Coefficient

- Models the log of the hazard as a function of the log of the baseline hazard (the hazard where all X=0, $h_0(t)$) and covariates (Xs).
- For a binary exposure *X*,
  - *c* = the log hazard ratio comparing hazard when *X* = 1 to *X* = 0
  - e^c = hazard ratio = relative hazard at all times *t* comparing hazard when *X* = 1 to *X* = 0
- Called a proportional hazards model because it assumes that the ratio of the hazards is constant across the time interval.
- Cox devised a method for estimating regression coefficients without needing to specify an intercept.

Proportional hazards

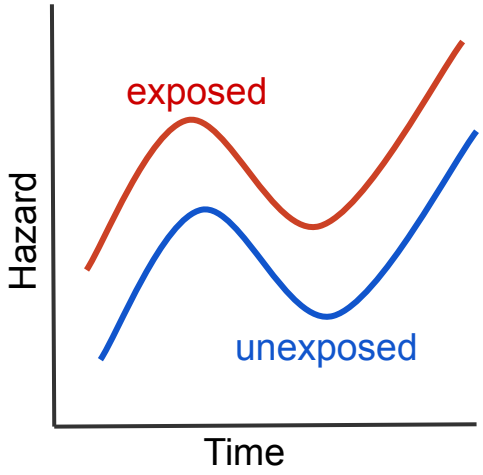
- Recall: Hazards cannot be directly calculated because they are defined for an infinitely small time interval

$$h(t) = \frac{P(\text{event in interval between } t \text{ and } [t + \Delta t] \mid \text{alive at } t)}{\Delta t}$$

- However, if the relative hazard can be assumed to remain constant over an interval it can be modeled without the need to estimate the actual hazard.
- The exposure is associated with a fixed relative increase in the hazard of the disease compared with the baseline hazard level.
  - At any time t, the hazard of the exposed ($h_1(t)$) is a multiple of the baseline hazard level ($h_0(t)$)
- Implication: the model provides one hazard ratio for the entire follow-up period.

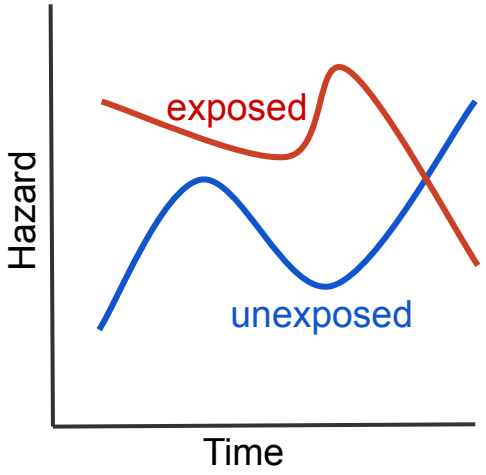
Proportional hazards

Proportional hazard



The model's proportionality assumption is valid.

Hazard that is not proportional



The model needs to stratify by periods of follow-up time so that the assumption is met within each stratum of time.

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$$

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

3. Next we fill in the coefficients for each term in the difference.

$$= (\ln h_0(t) + c\cdot(1)) - (\ln h_0(t) + c\cdot(0))$$

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

3. Next we fill in the coefficients for each term in the difference.

$$= (\ln h_0(t) + c\cdot(1)) - (\ln h_0(t) + c\cdot(0))$$

4. After simplifying, we see that c is equal to the log hazard ratio

$$= \ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = c$$

Obtaining the hazard ratio from a Cox proportional hazards model

1. Start with the Cox model specification

$$\ln\big(h(t|X=x)\big) = \ln h_0(t) + cx$$

2. We use the notation from the model to write the formula for the hazard ratio

$$\ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = \ln\big(h(t|X=1)\big) - \ln\big(h(t|X=0)\big)$$

3. Next we fill in the coefficients for each term in the difference.

$$= (\ln h_0(t) + c\cdot(1)) - (\ln h_0(t) + c\cdot(0))$$

4. After simplifying, we see that c is equal to the log hazard ratio

$$= \ln\left(\frac{h(t|X=1)}{h(t|X=0)}\right) = c$$

5. Exponentiating, e^c is equal to the hazard ratio

$$\frac{h(t|X=1)}{h(t|X=0)} = e^c$$

Example of Cox proportional hazards model

$$\ln(\text{hazard ratio}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Table 7-21 Results of a Cox proportional regression analysis of binary and continuous predictors of coronary heart disease (CHD) incidence in the Washington County cohort of the Atherosclerosis Risk in Communities (ARIC) Study, ages 45-64 years at baseline, 1987-1994

Variable (unit)	Cox regression coefficient	Hazard ratio
Gender (male = 1, female = 0)	1.2569	3.52
Smoking (yes = 1, no = 0)	0.7045	2.02
Age (1 year)	0.0120	1.012
Systolic blood pressure (1 mm Hg)	0.0152	1.015
Serum cholesterol (1 mg/dL)	0.0067	1.007
Body mass index (1 kg/m^2)	0.0237	1.024

<u>Binary covariate</u>

Hazard ratio for CHD for males compared to females is $e^{1.2569} = 3.52$

<u>Continuous covariate</u>

Hazard ratio for CHD for a 1 mm Hg increase in systolic blood pressure is $e^{0.0152} = 1.015$

Summary of key points

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Time-to-event outcome**: Cox proportional hazards model
- **How to interpret the coefficients from commonly used regression models**
  - $\ln\left(h(t|X=x)\right) = \ln h_0(t) + cx$
  - The exponentiated coefficient *c* (e^c) is the hazard ratio comparing the hazard when x = 1 to x = 0
- **Articulate certain assumptions of these models**
  - Called a proportional hazards model because it assumes that the ratio of the hazards is constant across the time interval.

Berkeley School of Public Health

Models for matched case-control data

PHW250B

Epidemiologic analysis topics (this week)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses
    - Linear regression
    - Logistic regression
    - Log-linear regression
- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data
  - Matched data **This video**

Modeling approaches for each type of matching

- **Individual matching of cases and controls**
  - Conditional logistic regression
  - Matching variables are included in the model by including a variable that links the members of each pair to each other
- **Frequency matching**
  - Regular logistic regression
  - Matching variables are included in the model as covariates
- **Matching on person-time** (density sampled designs)
  - Conditional logistic regression
  - Matching variables are included in the model by including a variable that links the members of each pair to each other
  - Length of follow-up (person-time contribution) is included in the model

Recap: Odds ratio formula for matched pair data - case-control study

- When case and control both either exposed or unexposed we get no information about the exposure disease relation.
- The only information is in the discordant pairs.
- **Odds Ratio = B / C**
  - (See Jewell pg 261 for the derivation)
- **Intuition**: B and C are the pairs in which we have variation in the exposure – if no variation in exposure, cannot look at relation between exposure and disease
- This formula does not allow us to adjust for potential confounders

Concordant pairs
Discordant pairs

Table 16.3 *Exposure patterns in the four types of matched pairs*

(1)	D	$\bar{D}$	
E	1	1	2
$\bar{E}$	0	0	0
	1	1	

(2)	D	$\bar{D}$	
E	1	0	1
$\bar{E}$	0	1	1
	1	1	

(3)	D	$\bar{D}$	
E	0	1	1
$\bar{E}$	1	0	1
	1	1	

(4)	D	$\bar{D}$	
E	0	0	0
$\bar{E}$	1	1	2
	1	1	

Organization of matched pair data

		Control	
		E	$\bar{E}$
Case	E	A	B
	$\bar{E}$	C	D

N

Jewell. *Statistics for Epidemiology.* 2004.

Motivation for conditional logistic regression

- In a matched case-control study, when we need to adjust for potential confounders (which we usually do) that were not matched on, we could use stratification-based methods
  - However, if pairs do not share the confounder value, they do not contribute to the odds ratio estimate
  - This can result in a loss of precision
- In this situation, a regression modeling approach is desirable.
- It also allows you to adjust for continuous confounders — stratification approaches require you to make these variables binary or categorical.
- Conditional logistic regression is analogous to logistic regression, but it takes into account the individual matching of cases and controls

Conditional logistic regression

$$\text{logit}(E(Y|X=x, I=i)) = \underbrace{\beta_0^*}_{\text{Coefficients}} + \beta_1 x \qquad \beta_0^* = \beta_0 + \beta_i$$

Indicators for matching stratum

- For a binary exposure X, β_1 = the difference in the log odds of the outcome Y when $X = 1$ and when $X = 0$
  - $e^{\beta 1}$ is the odds ratio comparing $X = 1$ to $X = 0$
- Intercept β_0* is equal to $\beta_0 + \beta_i$, the sum of the overall intercept and each stratum-specific intercept.
- These odds ratios adjust for potential confounders as well as matching variables if the additional potential confounders have been included in the regression model.
- This is analogous to including an indicator variable for each stratum of matching factors – in pair matching this would mean including an indicator variable for each pair.
- Assumes no clustering or auto-correlation

Summary of key points

- **Identify which type of model is used depending on the type of dependent and independent variable**
  - **Individually matched case-control study with binary outcome**: conditional logistic regression
- **How to interpret the coefficients**
  - $\text{logit}(E(Y|X = x, I = i)) = \beta_0^* + \beta_1 x$
  - The exponentiated coefficient β_1 ($e^{\beta 1}$) is the odds ratio comparing x = 1 to x = 0
- **Articulate certain assumptions of these models**
  - No clustering or auto-correlation