Categorical Predictors with Two Levels and Inference with Regression

Chapter 6, Lab 4

OpenIntro Biostatistics

Topics

- Categorical predictors with two levels
- Inference with regression

This lab introduces the idea of using a categorical predictor variable (specifically, a categorical predictor with two levels) in regression and also discusses the extension of statistical inference to the regression context.

The material in this lab corresponds to Sections 6.3.3 and 6.4 of *OpenIntro Biostatistics*.

Introduction

Categorical predictors with two levels

Although the response variable in linear regression is necessarily numerical, the predictor variable may be either numerical or categorical. Simple linear regression only allows for categorical predictors with two levels; multiple linear regression is required to examine categorical predictors with more than two levels.

Fitting a simple linear regression model with a categorical predictor that has two levels is analogous to comparing the means of two groups, where the groups are defined by the categorical variable. The equation of the regression line has intercept b_0, which equals the mean of one of the groups, and slope b_1, which equals the difference in means between the two groups.[1]

Inference with regression

When conducting inference in a regression context, observed data (x_i, y_i) used for fitting a regression line are assumed to have been randomly sampled from a population where the explanatory variable X and response variable Y follow a population model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\epsilon \sim N(0, \sigma)$. Under this assumption, the slope and intercept of the regression line, b_0 and b_1, are estimates of the population parameters β_0 and β_1.

Hypothesis tests and confidence intervals for regression population parameters have the same basic form as tests and intervals about population means. Inference is usually done about the slope, β_1. Under the null hypothesis, the variables X and Y are not associated; $H_0 : \beta_1 = 0$. Under the alternative hypothesis, the variables X and Y are associated; $H_1 : \beta_1 \neq 0$.

[1] The group for which b_0 is the mean is usually referred as the *baseline* group or *reference* group.

Categorical predictors with two levels

Inference with regression