Categorical Predictors with Several Levels and Inference in Regression

Chapter 7, Lab 3

OpenIntro Biostatistics

Topics

- Categorical predictors with several levels
- Inference in multiple regression

This lab expands on the topics introduced in Chapter 6, Lab 4 (Categorical Predictors with Two Levels and Inference in Regression) by introducing categorical predictors with more than two levels and generalizing inference in regression to the setting where there are several slope parameters.

The material in this lab corresponds to Sections 7.4 - 7.6 in *OpenIntro Biostatistics*.

Introduction

Categorical predictors with several levels

Fitting a regression model with a categorical predictor that has several levels is analogous to comparing the means of several 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 slopes $b_1, b_2, \dots, b_{p+1}$, where $p+1$ equals the number of groups and each slope b_k for $k = 1, 2, \dots, p+1$ equals the difference in means between the reference group and group k.

Inference in multiple regression

The observed data $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$ for $i = 1, 2, \dots, n$ cases are assumed to have been randomly sampled from a population where the response variable Y and p explanatory variables $X_1, X_2, \dots, X_p$ follow a population model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma)$. Under this assumption, the intercept and slopes of the regression line, b_0 and $b_1, b_2, \dots, b_p$, are estimates of the population parameters β_0 and $\beta_1, \beta_2, \dots, \beta_p$.

In multiple regression, the coefficient β_j of a predictor X_j denotes the change in the response variable Y associated with a one unit change in X_j when the values of the other predictors are held constant.

Hypothesis tests and confidence intervals for regression population parameters have the same basic structure as tests and intervals about population means. Inference is usually done about the slope parameters, $\beta_1, \beta_2, \dots, \beta_p$.

The F-statistic is used in an overall test of the model to assess whether the predictors in the model, considered as a group, are associated with the response.

Categorical predictors with several levels

Inference in regression

The t-statistic for a null hypothesis $H_0 : \beta_k = \beta_k^0$ has degrees of freedom $df = n - p - 1$, where n is the number of cases and p is the number of predictors in the model. The value β_k^0 equals 0 when the null hypothesis is one of no association.

$$t = \frac{b_k - \beta_k^0}{\text{s.e.}(b_1)} = \frac{b_k}{\text{s.e.}(b_1)}$$

A 95% confidence interval for β_k has the following formula, where $t^\star$ is the point on a t-distribution with $n - p - 1$ degrees of freedom and $\alpha/2$ area to the right.

$$b_k \pm \left(t^\star \times \text{s.e.}(b_k)\right)$$

The F-statistic in multiple regression is used to test hypotheses similar to those in ANOVA. The null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ is tested against the alternative that at least one of the slope coefficients is not 0. A significant p-value for the F-statistic is evidence that the predictor variables in the model, when considered as a group, are associated with the response variable.