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 discussing 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 and 7.9 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 $b_1, b_2, \ldots, b_{p+1}$, where $p+1$ equals the number of groups and each slope b_k for $k = 1, 2, \ldots, p+1$ equals the difference in means between the reference group and group k.

Inference in multiple regression

The observed data $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, 2, \ldots, n$ cases are assumed to have been randomly sampled from a population where the response variable Y and p explanatory variables $X_1, X_2, \ldots, X_p$ 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 $b_1, b_2, \ldots, b_p$, are estimates of the population parameters β_0 and $\beta_1, \beta_2, \ldots, \beta_p$.

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, $\beta_1, \beta_2, \ldots, \beta_p$.

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

The variable `Education` in the PREVEND data indicates the highest level of education that an individual completed in the Dutch educational system: primary school, lower secondary school, higher secondary education, or university education. This following questions step through exploring the association between RFFT score (`RFFT`) and educational level (`Education`) in `prevend.samp`, a random sample of $n = 500$ individuals from the PREVEND data.

1. Load `prevend.samp` and convert `Education` to a factor variable. The variable currently takes on values of either `0`, `1`, `2`, or `3`, where `0` denotes at most a primary school education, `1` a lower secondary school education, `2` a higher secondary education, and `3` a university education.
2. Identify how many individuals are in each level of `Education`.
3. Create a plot that shows the association between RFFT score and educational level. Describe what you see.
4. Calculate mean RFFT score for each educational attainment group.
5. Fit a linear regression model relating RFFT score and educational attainment.
    a) Write the equation of the least-squares line in terms of the variable names (e.g., *RFFT*).
    b) Based on part a), solve for the four possible values of $\widehat{RFFT}$ and interpret the values.
    c) Confirm that the numbers obtained in part b) match those from Question 4.
    d) Using a residual plot and a Q-Q plot, check the assumptions for linear regression. It is reasonable to assume that these observations are independent. Why is it not necessary to check the linearity assumption for the predictors in this model?

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.

6. Carry out inference based on the linear model in Question 5.
    a) Identify the t-statistics and p-values for each slope coefficient in the model. Interpret the p-values in the context of the data.
    b) Calculate and interpret the 95% confidence interval for the slope coefficient of X_3, university education.
    c) From the F-statistic, determine whether there is evidence of a significant association between RFFT score and educational level.
7. Conduct ANOVA to compare mean RFFT score among the four educational levels. Compare the results of inference based on the linear model to those based on ANOVA. For comparison purposes, leave the p-values uncorrected.
8. Suppose that the linear model in Question 5 had been fit with the original version of `Education` that had not been converted to a factor.
    a) Load the `prevend.samp` data to return `Education` to its original coding as an integer vector.
    b) Fit a linear model predicting RFFT from educational level without converting `Education` to a factor.
        i. Interpret the slope coefficient of the model.
        ii. What does this model imply about the change in mean RFFT between groups? Explain why this model is flawed.

Reanalyzing the PREVEND data

The following questions return to examining the association between cognitive decline and statin use, after adjusting for potential confounders.

In addition to age, there are two natural candidates for potential confounders: educational level and presence of cardiovascular disease (CVD). Individuals with more education tend to have higher incomes and consequently, better access to health care and medication; also, individuals with more education may be more comfortable with assessments like the RFFT. Individuals with cardiovascular disease are often prescribed statins to lower cholesterol; cardiovascular disease can lead to vascular dementia and cognitive decline.

9. Fit the multiple regression model relating RFFT (`RFFT`) with statin use (`Statin`), adjusting for the potential confounders age (`Age`), educational level (`Education`), and presence of CVD (`CVD`). The variable `CVD` is coded `0` if CVD is absent and `1` if CVD is present.
10. Based on the model from Question 9, summarize the evidence for an association between statin use and decreased cognitive function.
11. Evaluate the fit of the model.
    a) Assess the assumptions for linear regression.
    b) Comment on the R^2 and adjusted R^2 of the model. The adjusted R^2 from the model including only age as a potential confounder is 0.282.