Categorical Predictors with Two Levels and Inference in Regression

Chapter 6, Lab 3

OpenIntro Biostatistics

Topics

- Categorical predictors with two levels
- Inference in 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; examining categorical predictors with more than two levels requires multiple linear regression.

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 in 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 intercept and slope 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 structure 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

Obesity is known to be a leading risk factor for diabetes. The following questions step through exploring the association between BMI (`BMI`) and presence of diabetes (`DM`) in a random sample of $n = 500$ individuals from the PREVEND data.

1. Run the code chunk shown in the template to load `prevend.samp`, the random sample of 500 individuals from the PREVEND data used in the previous labs in this chapter.
2. Examine the variable `DM` and identify how many individuals in `prevend.samp` have diabetes. Presence of diabetes is coded as `1` and absence is coded as `0`.
3. Create plots that show the association between BMI and presence of diabetes.
   a) Create a boxplot that shows the association between BMI and presence of diabetes.
   b) Create a scatterplot of BMI versus presence of diabetes and plot the least-squares line.
   c) Based on the plots, comment briefly on the nature of the association.
4. In the Chapter 1 Lab Notes, the **factor** data structure was introduced. Factors are ideal for storing categorical data. In a factor variable, the levels of the variable are displayed as characters (such as "Female" and "Male") while the data remain stored as integer values (such as `0` and `1`); each level of the variable is associated with a specific integer value.
   a) Run the following code chunk to create the factor variable `DM.factor` from the integer vector `DM`. Note that while the `DM` variable is part of the `prevend.samp` dataframe, the variable `DM.factor` is not.

```
#create DM.factor
DM.factor = factor(prevend.samp$DM, levels = c(0, 1),
                                labels = c("Absent", "Present"))
```

   b) Run the `summary()` function on both `DM.factor` and `DM`. Compare the output and comment on which one has interpretive meaning.
   c) Run the following code chunk to overwrite the `DM` variable (in `prevend.samp`) with the factor variable `DM.factor`. Confirm that the overwrite was successful by running `summary()` on `DM`.

```
#overwrite DM with DM.factor
prevend.samp$DM <- DM.factor
```

5. Calculate mean BMI for diabetic individuals and individuals without diabetes.
6. Use a linear regression model to relate BMI and diabetes presence.
   a) Using a residual plot and Q-Q plot, check the assumptions for linear regression. It is reasonable to assume that these observations are independent.
   b) Write the equation of the least-squares line in terms of the variable names (e.g., *BMI*).
   c) Based on part b), solve for the two possible values of $\widehat{BMI}$ and interpret the values.
   d) Confirm that the numbers obtained in part c) match those from Question 5.

Inference in regression

Inference in a regression context is usually for the slope parameter, β_1.

The null hypothesis in regression is most commonly a hypothesis of 'no association', similar to how the null hypothesis when testing for a difference of means is often one of 'no difference'. When two variables are not associated, plotting them against each other results in a cloud of points with no apparent trend; in this setting, the slope of a least-squares line equals 0.

Thus, the hypotheses in regression can be written as:

- $H_0 : \beta_1 = 0$, the X and Y variables are not associated
- $H_A : \beta_1 \neq 0$, the X and Y variables are associated

The t-statistic for a null hypothesis $H_0 : \beta_1 = \beta_1^0$ has degrees of freedom $df = n - 2$, where n is the number of ordered pairs in the dataset. The value β_1^0 equals 0 when the null hypothesis is one of no association.

$$t = \frac{b_1 - \beta_1^0}{\text{s.e.}(b_1)} = \frac{b_1}{\text{s.e.}(b_1)}$$

A 95% confidence interval for β_1 has the following formula, where $t^\star$ is the point on a t-distribution with $n - 2$ degrees of freedom and $\alpha/2$ area to the right.

$$b_1 \pm \left(t^\star \times \text{s.e.}(b_1)\right)$$

The formulas for calculating the standard error of b_1 (s.e.(b_1)) are in Section 6.4 of *OpenIntro Biostatistics*. In practice, statistical software like R is used to obtain t-statistics and p-values for linear models.

7. Carry out inference based on the linear model from Question 6.

    a) Conduct a formal hypothesis test of no association between BMI and diabetes presence using `prevend.samp`, at the $\alpha = 0.05$ significance level.

        i. State the hypotheses.

        ii. Identify the relevant t-statistic and p-value from the output of the `summary(lm( ))` function.

        iii. State a conclusion in the context of the data.

    b) Calculate and interpret the 95% confidence interval for the slope coefficient of the model.

8. Use `t.test( )` to conduct a t-test for the difference in mean BMI between diabetic and non-diabetic individuals. Compare the results of inference based on the linear model to those based on a two-group test.

The following questions return to the investigation of RFFT score as a main response variable of interest in the PREVEND data.

9. The previous labs in this chapter have focused on exploring the association between RFFT score and age. The linear model with RFFT score as a response variable and age as a predictor was shown to reasonably satisfy the assumptions of linear regression.

a) Briefly discuss whether there is significant evidence of an association between RFFT score and age; be sure to report the relevant numerical evidence.

b) Compute and interpret the 99% confidence interval for the model slope.

10. Use a linear regression model to relate BMI and gender (`Gender`).

a) Convert `Gender` to a factor variable. In the original variable, males are coded as `0` and females are coded as `1`.

b) Fit the model and interpret the model intercept and slope.

c) Evaluate whether gender is a significant predictor of BMI.