Evaluating Model Fit

Chapter 7, Lab 2

OpenIntro Biostatistics

Topics

- Checking model assumptions
- Using R^2 and adjusted R^2

Methods for evaluating the fit of a multiple regression model are similar to those for a simple regression model. Residual plots become essential tools for checking modeling assumptions since it is not possible to make a two-dimensional plot of a response variable against several predictors simultaneously. This lab discusses the use of residual plots to check assumptions for multiple linear regression and introduces adjusted R^2.

The material in this lab corresponds to Section 7.3 of *OpenIntro Biostatistics*.

Introduction

Assumptions for multiple linear regression

1. *Linearity*: For each predictor variable x_i, change in the predictor is linearly related to change in the response variable when the value of all other predictors is held constant.
2. *Constant variability*: The residuals have approximately constant variance.
3. *Independent observations*: Each set of observations $(y, x_1, x_2, \ldots, x_p)$ is independent.
4. *Approximate normality of residuals*: The residuals are approximately normally distributed.

Plots for checking assumptions

The linearity assumption is assessed with respect to each predictor variable. For each predictor, examine a residual plot in which the values of the predictor variable are on the x-axis and the model residuals are on the y-axis. Any patterns or curvature are indicative of non-linearity, as in the residual plot for simple linear regression.

Since each case in a dataset has one residual value and one predicted (i.e., fitted) value, regardless of the number of predictors in the model, the constant variance assumption can still be assessed with the same method as for simple linear regression: examining a scatterplot of predicted values on the x-axis and residual values on the y-axis.

Normal probability plots can be used to check the normality of the residuals.

R^2 and adjusted R^2

Adding a variable to a regression model always increases the value of R^2. The **adjusted** R^2 imposes a penalty for including predictors that do not contribute much towards explaining observed variation in the response variable.

RFFT, statin use, and age in the `prevend` data

The questions in this section use data from `prevend.samp`, a random sample of $n = 500$ individuals from the `prevend` dataset. Run the code shown in the template to load `prevend.samp` from the `oibiostat` package and convert the statin use variable into a factor.

1. Fit a multiple regression model predicting RFFT score from statin use and age. Check the assumptions for multiple linear regression.
    a) Assess linearity with respect to age using a scatterplot with residual values on the y-axis and values of age on the x-axis. Is it necessary to assess linearity with respect to statin use?
    b) Assess whether the residuals have approximately constant variance.
    c) Is it reasonable to assume that each set of observations is independent of the others?
    d) Assess whether the residuals are approximately normally distributed.
2. How well does the model explain the variability in observed RFFT score?

The **adjusted** R^2 is computed as

$$R^2_{adj} = 1 - \left(\frac{\text{Var}(e_i)}{\text{Var}(y_i)} \times \frac{n-1}{n-p-1} \right),$$

where n is the number of cases and p is the number of predictor variables.

3. Using the formula, calculate the adjusted R^2 for the multiple regression model predicting RFFT score from statin use and age.
4. In the previous lab, a multiple regression model was used to estimate the association between statin use and RFFT score while adjusting for age as a potential confounder. Suppose that instead, the goal was to build a model that effectively explains the observed variation in RFFT score; in other words, to build a predictive model for RFFT score. In such a setting, there is no primary predictor of interest.

    The adjusted R^2 is useful as a statistic for comparing models to select a best predictive model. Model selection will be discussed in Chapter 7, Lab 5.

    While R^2 increases with the addition of any predictor to a model, adjusted R^2 only increases substantially when a 'useful' predictor is added; that is, a predictor that contributes to explaining observed variation in the response.

    a) Would you expect the adjusted R^2 for the multiple regression model to be very different from the adjusted R^2 for the simple regression model predicting RFFT score from age? Explain your answer.
    b) Would you expect the adjusted R^2 for the multiple regression model to be very different from the adjusted R^2 for the simple regression model predicting RFFT score from statin use? Explain your answer.

Expenditures, race, and age in the `dds.discr` data

Recall that Chapter 1 introduced a case study examining the evidence for ethnic discrimination in the amount of financial support offered by the State of California to individuals with developmental disabilities. Although an initial look at the data suggested an association between expenditures and ethnicity (specifically between Hispanics and White non-Hispanics), further analysis suggested that age is a confounding variable for the relationship.

The amount of financial support provided to individuals is stored as the `expenditures` variable. The age variable records age in years and the `ethnicity` variable records ethnicity. The data in `dds.discr` represent a random sample of 1,000 individuals who receive financial support from the California Department of Developmental Services (out of a total population of 250,000).

5. Run the following code to load the `dds.discr` data from the `oibiostat` package and subset the data to include only observations from Hispanic and White non-Hispanic consumers.

```
#load the data
data("dds.discr")

#subset the data
dds.subset = dds.discr[dds.discr$ethnicity == "Hispanic" |
                       dds.discr$ethnicity == "White not Hispanic", ]

#drop unused factor levels
dds.subset$ethnicity = droplevels(dds.subset$ethnicity)
```

6. Fit a multiple regression model predicting expenditures from ethnicity and age using the data in `dds.subset` and write the equation of the linear model.

7. Evaluate the model fit.

    a) Assess linearity with respect to age.

    b) Assess whether the residuals have approximately constant variance.

    c) Is it reasonable to assume that each set of observations is independent of the others?

    d) Assess whether the residuals are approximately normally distributed.

    e) Overall, does a linear model seem appropriate for modeling the relationship between expenditures, ethnicity, and age?