Lab Notes

Unit 9

Statistics 102

Overview

1. Simple Logistic Regression
   - *OI Biostat* Section 9.xx
2. Multiple Logistic Regression
   - *OI Biostat* Section 9.xx

Lab 1 introduces the multiple regression model in the context of estimating an association between a response variable and primary predictor of interest while adjusting for possible confounding variables.

Lab 2 discusses the use of residual plots to check assumptions for multiple regression and introduces adjusted R^2.

Lab 1: Simple Logistic Regression

Working with Several Predictors

The **lm()** function is used to fit linear models. It has the following generic structure:

```
lm(y ~ x1 + x2, data)
```

where the first argument specifies the variables used in the model; in this example, the model regresses a response variable y against two explanatory variables x1 and x2. Additional predictor variables can be added to the model formula with the + symbol.

The following example shows fitting a linear model that predicts BMI from age (in years) and gender using data from `nhanes.samp.adult.500`, a sample of individuals 21 years of age or older from the NHANES data.

```
#load the data
library(oibiostat)
data("nhanes.samp.adult.500")

#fitting linear model
lm(BMI ~ Age + Gender, data = nhanes.samp.adult.500)
```

```
##
## Call:
## lm(formula = BMI ~ Age + Gender, data = nhanes.samp.adult.500)
##
## Coefficients:
## (Intercept)          Age   Gendermale
##    28.80865      0.02064     -0.95709
```

Letting R do the Work: Predicted Values

The `predict()` function can be used to evaluate the regression equation for specific x-values, or in other words, to calculate $\hat{y}$ values for values of x that were not necessarily observed. To use `predict()` in this way, specify the x-values according to the following generic syntax:

```
predict(object, newdata = data.frame( ))
```

where `object` is the name of the fitted model, and the name of the predictor variable and value at which to evaluate the equation are specified within `newdata = data.frame()`.

In a model with several variables, values for all variables in the model must be specified to calculate a prediction.

The following example shows calculating $\widehat{BMI}$ for a male individual 60 years of age using the model regressing BMI on age and gender in `nhanes.samp.adult.500`, then checking the result by explicitly solving the regression equation.

```
#BMI ~ Age + Gender in nhanes.samp.adult.500
model.BMIvsAgeGender = lm(BMI ~ Age + Gender, data = nhanes.samp.adult.500)
predict(model.BMIvsAgeGender, newdata = data.frame(Age = 60, Gender = "male"))
```

```
##     1
## 29.09
```

```
#confirm answer from solving 28.81 + 0.02(60) - 0.95(1)
coef(model.BMIvsAgeGender)[1] + coef(model.BMIvsAgeGender)[2]*60 +
  coef(model.BMIvsAgeGender)[3]*1
```

```
## (Intercept)
##       29.09
```

Lab 1: Multiple Logistic Regression