Class 7: Chapter 8

EPH 705

Min Lu

Division of Biostatistics
University of Miami

Spring 2017

Overview

1. Object:
   - Hosmer –Lemeshow test
   - Plot and Understand Model Specification
2. R Exercise
   - Plot Estimated Curve with Observed Data
   - In class exercise
   - Take home exercise

Hosmer–Lemeshow test

The Hosmer–Lemeshow test is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are called well calibrated.

Hosmer–Lemeshow test

The Hosmer–Lemeshow test statistic is given by

$$
\begin{aligned}
H &= \sum_{g=1}^{G} \frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}} \\
&= \sum_{g=1}^{G} \frac{(O_{1g} - E_{1g})^2}{N_g \pi_g} + \frac{(N_g - O_{1g} - (N_g - E_{1g}))^2}{N_g(1 - \pi_g)} \\
&= \sum_{g=1}^{G} \frac{(O_{1g} - E_{1g})^2}{N_g \pi_g (1 - \pi_g)}
\end{aligned}
$$

Here O_{1g}, E_{1g}, O_{0g}, E_{0g}, N_g, and π_g denote the observed $Y = 1$ events, expected $Y = 1$ events, observed $Y = 0$ events, expected $Y = 0$ events, total observations, predicted risk for the gth risk decile group, and G is the number of groups. The test statistic asymptotically follows a χ^2 distribution with G - 2 degrees of freedom. The number of risk groups may be adjusted depending on how many fitted risks are determined by the model. This helps to avoid singular decile groups.

Western Collaborative Group Study data

Data:The Western Collaborative Group Study (WCGS), a prospective cohort studye, recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured.

variable name	discreption
id	Subject ID:
age0	Age: age in years
height0	Height: height in inches
weight0	Weight: weight in pounds
sbp0	Systolic blood pressure: mm Hg
dbp0	Diastolic blood pressure: mm Hg
chol0	Cholesterol: mg/100 ml
behpat0	Behavior pattern:
ncigs0	Smoking: Cigarettes/day
dibpat0	Dichotomous behavior pattern: 0 = Type B; 1 = Type A
chd69	Coronary heart disease event: 0 = none; 1 = yes
typechd	to be done
time169	Observation (follow up) time: Days
arcus0	Corneal arcus: 0 = none; 1 = yes

Hosmer-Lemeshow Tests for Logistic Regression Models

```
library("generalhoslem")
wcgs <- read.csv("wcgs.csv")[, -1]
## Binary model
mod1 <- glm(chd69 ~ age0 + dibpat0, data = wcgs, family = binomial(link = "logit"))
logitgof(wcgs$chd69, fitted(mod1), g = 10)
```

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  wcgs$chd69, fitted(mod1)
## X-squared = 8.1057, df = 8, p-value = 0.4232
```

Model Specification

Outcome $Y_i \sim \text{Bin}(n_i, p_i)$, for $i = 1, \ldots, n$

$$\Pr(Y_i = y) = \binom{n_i}{y} {p_i}^y (1 - p_i)^{n_i - y}$$

Logistic regression for binomial outcome

$$\Pr(Y_i = y \mid \mathbf{X}_i) = \binom{n_i}{y} \left(\frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}}\right)^y \left(1 - \frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}}\right)^{n_i - y}.$$

And the logistic function can now be written as: $F(x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$.

For a continuous independent variable x the odds ratio can be defined as:

$$\text{OR} = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \frac{\left(\frac{F(x+1)}{1-F(x+1)}\right)}{\left(\frac{F(x)}{1-F(x)}\right)} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

This exponential relationship provides an interpretation for β_1 : The odds multiply by e^{β_1} for every 1-unit increase in x.

Model Specification Example

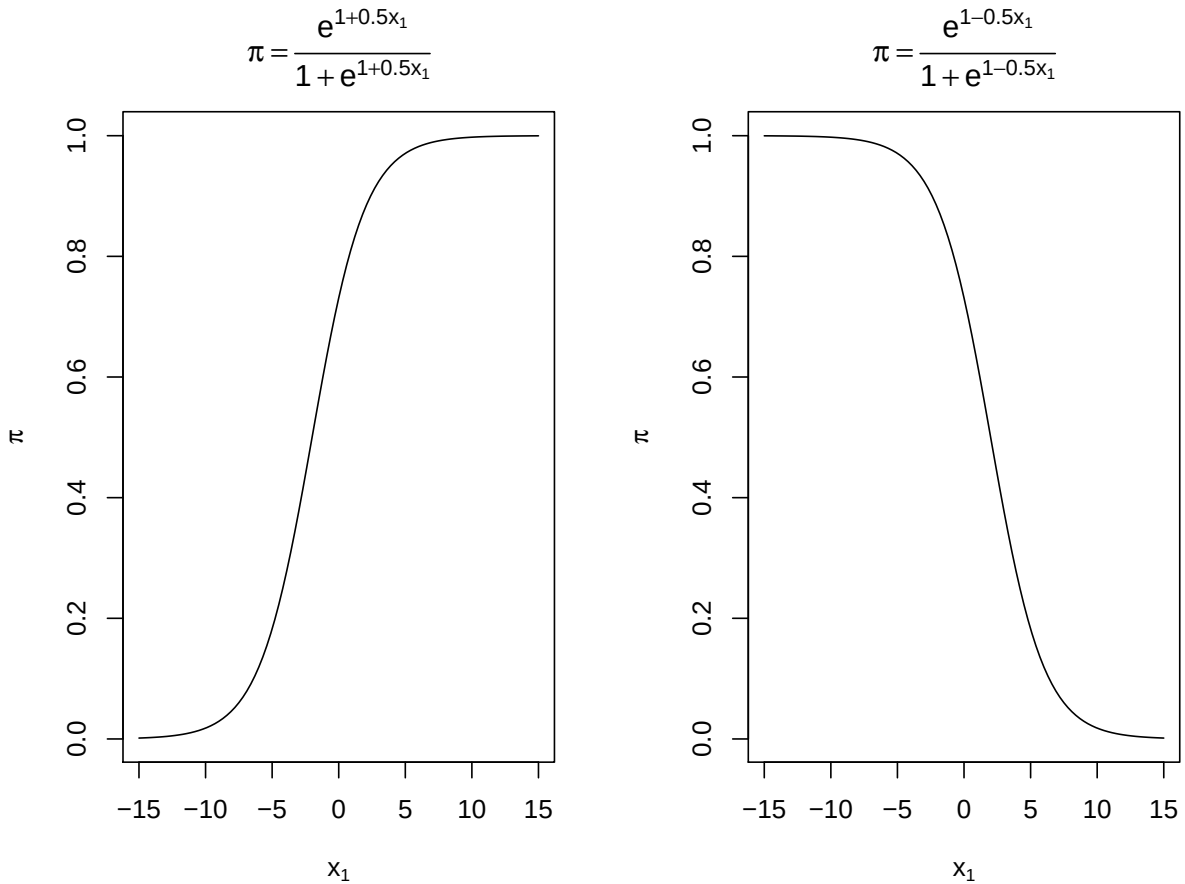

Figure 1.1

Model Specification Example

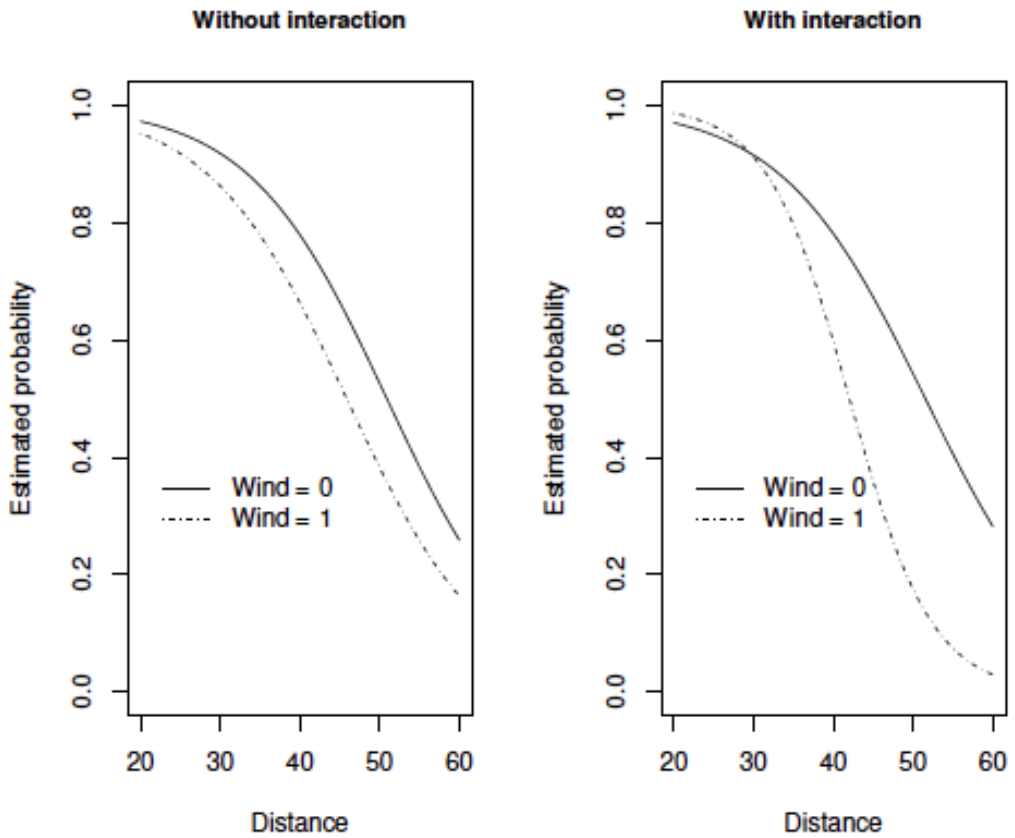

Figure 1.2

Estimated Curve in Logistic Regression without Interaction

```
mod.fit <- glm(chd69 ~ age0, data = wcgs, family = binomial(link = "logit"))
# Find observed proportion of Coronary heart disease at each age
w <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = sum)
n <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = length)
w.n <- data.frame(Age = w$age0, With.Coronary.H.Disease = w$chd69, Total = n$chd69, proportion = round(w$chd69/n$chd69,
    4))
head(w.n)
```

```
##   Age With.Coronary.H.Disease Total proportion
## 1  39                      19   266     0.0714
## 2  40                      12   277     0.0433
## 3  41                      12   233     0.0515
## 4  42                       5   222     0.0225
## 5  43                      13   215     0.0605
## 6  44                      13   235     0.0553
```

Estimated Curve with Wald Confidence Interval

```
source("ci_function.R")
plot(x = w$age0, y = w$chd69/n$chd69, xlab = "Age", ylab = "Probability of coronary hear disease",
     panel.first = grid(col = "gray", lty = "dotted"))
curve(expr = predict(object = mod.fit,  newdata=data.frame(age0 = x),
                     type = "response"), col = "red", add = TRUE,
      xlim = c(min(wcgs$age0), max(wcgs$age0)))
curve(expr = ci.pi(newdata = data.frame(age0 = x),
                   mod.fit.obj = mod.fit, alpha = 0.05)$lower, col = "blue",
      lty = "dotdash", add = TRUE, xlim = c(min(wcgs$age0), max(wcgs$age0)))
curve(expr = ci.pi(newdata = data.frame(age0 = x),
                   mod.fit.obj = mod.fit, alpha = 0.05)$upper, col = "blue",
      lty = "dotdash", add = TRUE, xlim = c(min(wcgs$age0), max(wcgs$age0)))
legend(x = 40, y = 0.15, legend = c("Logistic regression model", "95% individual C.I."),
       lty = c("solid", "dotdash"), col = c("red", "blue"), bty = "n")
```

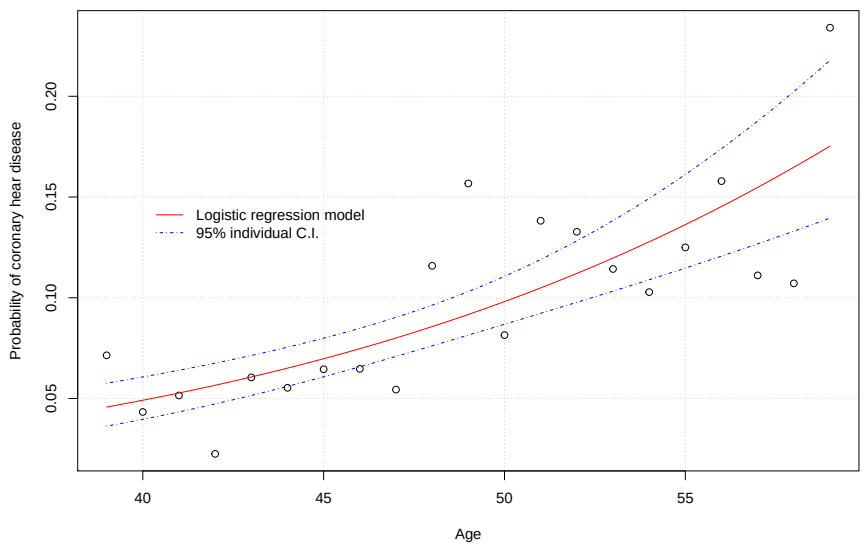

Figure 1.3

Estimated Curve in Logistic Regression without Interaction

```
wcgs <- read.csv("wcgs.csv")[, -1]
mod.fit <- glm(chd69 ~ age0, data = wcgs, family = binomial(link = "logit"))
# Find the observed proportion of Coronary heart disease at each age
w <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = sum)
n <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = length)
w.n <- data.frame(Age = w$age0, With.Coronary.H.Disease = w$chd69, Total = n$chd69, proportion = round(w$chd69/n$chd69,
    4))
head(w.n)
```

```
##   Age With.Coronary.H.Disease Total proportion
## 1  39                      19   266     0.0714
## 2  40                      12   277     0.0433
## 3  41                      12   233     0.0515
## 4  42                       5   222     0.0225
## 5  43                      13   215     0.0605
## 6  44                      13   235     0.0553
```

Figure 1.4

Estimated Curve in Logistic Regression without Interaction

```
wcgs <- read.csv("wcgs.csv")[, -1]
mod.fit <- glm(chd69 ~ age0, data = wcgs, family = binomial(link = "logit"))
# Find the observed proportion of Coronary heart disease at each age
w <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = sum)
n <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = length)
w.n <- data.frame(Age = w$age0, With.Coronary.H.Disease = w$chd69, Total = n$chd69, proportion = round(w$chd69/n$chd69,
    4))
head(w.n)
```

```
##   Age With.Coronary.H.Disease Total proportion
## 1  39                      19   266     0.0714
## 2  40                      12   277     0.0433
## 3  41                      12   233     0.0515
## 4  42                       5   222     0.0225
## 5  43                      13   215     0.0605
## 6  44                      13   235     0.0553
```

Figure 1.5

Estimated Curve in Logistic Regression without Interaction

```
wcgs <- read.csv("wcgs.csv")[, -1]
mod.fit <- glm(chd69 ~ age0, data = wcgs, family = binomial(link = "logit"))
# Find the observed proportion of Coronary heart disease at each age
w <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = sum)
n <- aggregate(formula = chd69 ~ age0, data = wcgs, FUN = length)
w.n <- data.frame(Age = w$age0, With.Coronary.H.Disease = w$chd69, Total = n$chd69, proportion = round(w$chd69/n$chd69,
    4))
head(w.n)
```

```
##   Age With.Coronary.H.Disease Total proportion
## 1  39                      19   266     0.0714
## 2  40                      12   277     0.0433
## 3  41                      12   233     0.0515
## 4  42                       5   222     0.0225
## 5  43                      13   215     0.0605
## 6  44                      13   235     0.0553
```

Figure 1.6

In class exercise

Use the wcgs data and plot Figure 1.6 but substitue variable dibpat0 with arcus0.

Take home exercise

Use the wcgs data, choose your own outcome and predictors, conduct a GLM with negative binomial family and make two plots: one from model with main effects only and the other from model with interaction. Save the plot in a PDF file.

Class over