Conditional Logistic Regression and GLMs

Lab 10

1 Logistic Regression for Retrospective and Matched Studies

1.1 Logistic Regression for Case-Control Data

In a cohort study, we randomly sample individuals regardless of their outcome/response variable. For example, we first randomly sample 1000 individuals, then we follow them up for 10 years and record their smoking status and whether they develop a lung cancer during that period.

Cancer	Non-smoker	Smoker
Yes	100	100
No	600	200

If we fit a logistic regression model $logit(p_i) = \alpha + \beta_1 I(smoker_i)$ *to study the relationship between lung cancer and smoking status, how do we estimate and interpret* α *and* β_1*?*

What's the probability of having lung cancer for a non-smoker?

Now, consider the same data assuming that they are actually collected from a retrospective, case-control study (we first select 200 cancer cases and 800 non-cases, then we do a survey to check how many are smokers/non-smokers).

Based on this design, consistent with the notes we had in class we fit a simple logistic model to study the relationship between lung cancer and smoking status. While we still estimate a two parameter model

$$\text{logit}(p_i) = \alpha^* + \beta_1 I(smoker_i),$$

under the case-control design the we know that the underlying true model is

$$\text{logit}(p_i) = \alpha + \log(\tau_1/\tau_0) + \beta_1 I(smoker_i)$$

where τ_1 and τ_0 are the 'sampling fractions' for cases and controls, respectively. We often don't know these values.

What changes about the estimates we get, or our interpretation of them?

Can we estimate the probability of having lung cancer for a non-smoker in a case-control study without additional information?

Which of the following statements is correct?
A. Only the estimated odds ratio from the cohort study is valid.
B. Only the estimated odds ratio from the case-control study is valid.
C. No matter which study design you choose, the odds ratio estimate is always valid.

2 Conditional Logistic Regression for Matched Data

The above univariate analysis was unadjusted, so the estimates ignored potential confounders such as income (continuous). As a result, the estimates above are likely biased.

If we still want to research this question, and are still interested in using a case-control approach, we could instead design a study and analysis that adjusts for income, using one of two approaches. Remember, our primary interest is to estimate the odds ratio of having lung cancer comparing smokers and non-smokers:

A. Collect a set of cases and a set of controls, and fit a multivariate logistic regression, namely, $\text{logit}(p) = \beta_0 + \beta_1 I(smoker) + \beta_2 * income$

B. Categorize income to many small groups, such as <20k,20-25k, 25-30k,...,>120k, find as many cases as possible, and then find 4 matched controls in the same category for each case. finally fit a conditional logistic model

$$\text{logit}(p_{ij}) = \alpha_i + \beta I_{ij}(smoker),$$

where i denotes the ith income category and j denotes the jth individual in that category.

(When comparing these two approaches, we could assume that the same data might arise under each design–in other words, method A and method B could yield the same data, but they would have come from different study designs, and so need to be analyzed using appropriate models.)

What would be the pros and cons of each approach?

*Now, assume that gender is also available in the dataset and we matched up both income and gender to estimate the odds ratio of lung cancer comparing smokers and non-smokers with a conditional logistic regression. Is that sufficient to adjust for the interaction between gender and income? If we are interested in the effect modification of gender to smoking effect as well, is it ok for us to add I(smoker)*gender into this conditional regression model?*

2.1 Conditional Logistic Regression Data Example

We will be using matched data from the Women's Health Study (WHS), a randomized trial that tested low dose aspirin (100 mg on alternate days) and vitamin E supplementation (600 IU natural source vitamin E on alternate days) for the primary prevention of cardiovascular disease and cancer in 39,876 women health professionals. This nested case-control study used baseline blood samples from 130 incident cases of cardiovascular disease (CVD) during follow-up, and matched each case to another participant who was free of CVD at the time of follow-up on age (within 2 years) and smoking status (current, past, or never). We will focus on systolic blood pressure.

[Quick note on the dataset: one control did not know her systolic blood pressure. SAS and Stata represent missing values as . while R represents them as NA. Furthermore, when subsetting datasets, Stata assumes missing values are very large positive numbers, whereas SAS assumes they are very small negative numbers, so a code for something like "keep all values greater than 140" would also include missing values in Stata, but not in SAS. Note the importance of accounting for missing values when we create indicator variables.]

Variables in the file whsmtch.dat are:

1. Id: 1-130 with a case and her control having the same value
2. Crpq: quartile of high-sensitivity C-reactive protein, based on the distribution in controls
3. Sbp: systolic blood pressure in mm Hg

4. Dm: diagnosis of diabetes mellitus (1=yes, 0=no)
5. Age: in years
6. Chdlq: quartile of total cholesterol to HDL cholesterol ratio
7. Caco: case-control indicator (1=case, 0=control)

Systolic blood pressure (SBP) was self-reported by women at baseline in one of 9 categories: <110, 110-119, 120-129, 130-139, 140-149, 150-159, 160-169, 170-179, ≥180 mm Hg. In this data set, we have used mid-point scoring, meaning that these categories appear in the data as their midpoint value. We will begin by reading in the data and creating a new binary exposure called hisbp, which we set equal to 1 if SBP $\geq$ 140 mm Hg and 0 otherwise.

```
# Set up the working directory and input our data
colnames(whs) = c("id", "crpq", "sbp", "dm", "age", "smok", "chdlq", "caco")

#subset to remove observation with missing value
whs = whs[whs$id != whs[whs$sbp == ".","id"],]

#convert SBP from factor to numeric
whs$sbp = as.numeric(levels(whs$sbp))[whs$sbp]

#create binary "High SBP" indicator
whs$high_sbp = as.numeric(whs$sbp >= 140)

head(whs)
```

```
##   id crpq sbp dm age smok chdlq caco high_sbp
## 1  1    3 175  0  82    2     3    0        1
## 2  1    4 155  0  84    2     4    1        1
## 3  2    2 115  0  78    2     4    1        0
## 4  2    2 135  0  78    2     3    0        0
## 5  3    2 135  0  78    2     4    0        0
## 6  3    4 155  0  77    2     1    1        1
```

2.1.1 Conditional Logistic Regression with a Single Binary Covariate

First, consider a conditional logistic model that includes only an indicator variable for high systolic blood pressure.

$$\text{logit}(p_{ij}) = \alpha_i + \beta_1 x_{ij}$$

Where i indexes the matched pairs and j indexes the subjects within each matched pair.

```
library(survival)
fit0 <- clogit(caco ~ high_sbp + strata(id), data=whs)
summary(fit0)
```

```
## Call:
## coxph(formula = Surv(rep(1, 258L), caco) ~ high_sbp + strata(id),
##     data = whs, method = "exact")
##
##   n= 258, number of events= 129
```

```
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## high_sbp 0.7259    2.0667   0.3145 2.308    0.021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## high_sbp     2.067     0.4839     1.116     3.828
##
## Rsquare= 0.022   (max possible= 0.5 )
## Likelihood ratio test= 5.68  on 1 df,   p=0.02
## Wald test            = 5.33  on 1 df,   p=0.02
## Score (logrank) test = 5.57  on 1 df,   p=0.02
```

Which variables does it adjust for?

Why isn't any coefficient estimate for a constant/intercept like α_i reported?

How do you interpret the results of this model? What can we say about the relationship between High SBP and risk of cardiovascular disease?

2.1.2 Conditional Logistic Regression with a Categorical Covariate

Now instead of a binary outcome for sbp, we form four ordinal categories of systolic pressure: [0,120), [120,140), [140,160), and [160,∞). We will now examine whether we can fit a model assuming an ordinal effect of sbp on the outcome, or if we need to have a more flexible categorical specification of the effect of sbp. The two candidate models can be written as follows:

$$\text{logit}(p_{ij}) = \alpha_i + \beta_1 I(sbp_{ij} \in [120, 140)) + \beta_2 I(sbp_{ij} \in [140, 160)) + \beta_3 I(sbp_{ij} \in [160, \infty)) \quad (1)$$

$$\text{logit}(p_{ij}) = \alpha_i + \beta_1 \times sbpcat_{ij} \quad (2)$$

Where *sbpcat* is 0 if sbp is below 120, 1 if sbp is between 120 and 140, 2 if sbp is between 140 and 160, and 3 if sbp is above 160.

Here are the model results for each of the two models:

```
#Create categorical SBP
whs$sbpcat = 1
whs$sbpcat = whs$sbpcat + (whs$sbp >119) + (whs$sbp >139) + (whs$sbp >159)


# Fit categorical model
fit1 = clogit(caco ~ as.factor(sbpcat) + strata(id), data = whs)
summary(fit1)
```

```
## Call:
## coxph(formula = Surv(rep(1, 258L), caco) ~ as.factor(sbpcat) +
##     strata(id), data = whs, method = "exact")
##
##   n= 258, number of events= 129
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## as.factor(sbpcat)2 0.4171    1.5176   0.3014 1.384   0.1663
## as.factor(sbpcat)3 0.9780    2.6592   0.3921 2.494   0.0126 *
## as.factor(sbpcat)4 1.1592    3.1875   0.7538 1.538   0.1241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                     exp(coef) exp(-coef) lower .95 upper .95
## as.factor(sbpcat)2     1.518     0.6589    0.8407     2.740
## as.factor(sbpcat)3     2.659     0.3761    1.2331     5.735
## as.factor(sbpcat)4     3.187     0.3137    0.7275    13.966
##
## Rsquare= 0.029   (max possible= 0.5 )
## Likelihood ratio test= 7.69  on 3 df,   p=0.05
## Wald test            = 7.13  on 3 df,   p=0.07
## Score (logrank) test = 7.5  on 3 df,   p=0.06
```

```
# Fit ordinal model
fit2 = clogit(caco ~ sbpcat + strata(id), data = whs)
summary(fit2)
```

```
## Call:
## coxph(formula = Surv(rep(1, 258L), caco) ~ sbpcat + strata(id),
##     data = whs, method = "exact")
##
##   n= 258, number of events= 129
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## sbpcat 0.4541    1.5748   0.1723 2.636   0.0084 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## sbpcat     1.575      0.635     1.123     2.207
##
## Rsquare= 0.029   (max possible= 0.5 )
## Likelihood ratio test= 7.51  on 1 df,   p=0.006
## Wald test            = 6.95  on 1 df,   p=0.008
## Score (logrank) test = 7.32  on 1 df,   p=0.007
```

```
#run Likelihood ratio test comparing categorical vs. ordinal
anova(fit1,fit2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(rep(1, 258L), caco)
##  Model 1: ~ as.factor(sbpcat) + strata(id)
##  Model 2: ~ sbpcat + strata(id)
##    loglik  Chisq Df P(>|Chi|)
## 1 -85.573
## 2 -85.659 0.1728  2    0.9172
```

State the null and alternative hypotheses for a likelihood ratio test comparing these two specifications. Run the test and state your conclusions.

Based on our preferred model we've just selected, how would we interpret these results?

3 Generalized Linear Models

Generalized Linear Models are a framework we will use throughout the rest of the semester to build regression models for outcomes that follow specific distributions. Formally, GLMs specify a parametric statistical model for the conditional distribution of some response Y_i given a $p+1$-vector of covariates $\mathbf{X}_i = (1, X_1, \ldots, X_p)^T$. (Don't be alarmed by the linear algebra notation, we will clarify this. For now, just know that this represents all of the different covariates we are using to model our outcome.)

GLMs are defined by

$$g(E[Y_i|\mathbf{X}_i]) = \mathbf{X}_i^T\boldsymbol{\beta}$$

and are governed by 3 components:

1. a probability distribution for the response, $Y \sim f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$
2. a linear predictor, $\mathbf{X}_i^T\boldsymbol{\beta}$
3. a link function, $g(\cdot)$

The one piece that might surprise you is the specific requirement for the probability distribution. Why that form specifically?

It turns out that GLMs need responses whose distribution belongs to what are called *exponential dispersion families*. A distribution belongs to the exponential dispersion family if its pdf/pmf can be written in the above form, where θ is called the canonical parameter, and ϕ the dispersion parameter. It may not look it, but many familiar probability distributions follow this form!

3.1 Logistic Regression as a GLM

In fact, the logistic regression we've already seen is a great example of a GLM.

Wait, are we sure that the Bernoulli distribution we model in logistic regression is an exponential dispersion family?

Now, confident that our outcome follows a compatible distribution, we turn to the specification of our model:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_1 x_{ip}$$

where $p_i = P(Y_i = 1|\mathbf{X}_i) = E[Y_i|\mathbf{X}_i]$.

What is the linear predictor?

What is the link function in logistic regression?

And that's it! We've just shown that logistic regression is an example of a GLM.

3.2 Linear Regression as a GLM

We can even look back at linear regression and show that it is an example of a GLM! It's more good practice for identifying all of the components of a GLM, and unlike in the above example where a single θ parameter was all we needed, here we'll find that we will use the dispersion parameter ϕ.

To refresh, the setting of linear regression is

$$E[Y_i|\mathbf{X}_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_1 X_{ip}$$

where each observation is normally distributed given $\mathbf{X}$, that is

$$Y|\mathbf{X} \sim f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Can we verify that this distribution is an exponential dispersion family?

Now, the other two components:

- Linear predictor: Same as before, $\mathbf{X}_i^T\boldsymbol{\beta} = \beta_0 + \sum_{i=1}^{p} \beta_i X_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Link function: this time, it is just what's called the 'identity' function: $g(E[Y_i|\mathbf{X}_i]) = E[Y_i|\mathbf{X}_i]$.

And that's it! We've characterized linear regression as a special case of a GLM.

3.3 Properties of Exponential Dispersion Families

For exponential dispersion families, the two following can be shown:

$$E(Y) = b'(\theta) = \mu$$
$$Var(Y) = \phi b''(\theta)$$

(Calculus is not a necessary prerequisite of this course, but just understanding that there are these deeper connections can help us to appreciate why we go to the effort to write out the distributions in this special form.)

We will show these properties using the simple example of the Exponential distribution:

$$f_Y(y; \lambda) = \lambda \exp(-\lambda y),\ y \in [0, \infty)$$

Recall, for this distribution, we can show from standard probability that $E(Y) = \frac{1}{\lambda}$ and $Var(Y) = \frac{1}{\lambda^2}$, so those are our points of comparison.

Show that this distribution is part of the exponential dispersion family, and calculate $E(Y)$ and $Var(Y)$ according to the formulas above.