Logistic Regression Part 2

Adam J Sullivan, PhD

02/28/2018

Logistic Model Building

Confounding and Effect Modification

- We can also address confounding and effect modification in R.
- To explore this more we will consider a new data example with a subset of data from the Phsyicians Health Study.

Physician Health Study Variables

NAME	DESCRIPTION
age	Age in years at time of Randomization
asa	0 - placebo, 1 - aspirin
bmi	Body Mass Index (kg/m^2)
hypert	1 - Hypertensive at baseline, 0 - Not
alcohol	0 - less than monthly, 1 - monthly to less than daily, 2 - daily consumption
dm	0 = No diabetes Mellitus, 1 - diabetes Mellitus
sbp	Systolic BP (mmHg)

Physician Health Study Variables

NAME	DESCRIPTION
exer	0 - No regular, 1 - Sweat at least once per week
csmoke	0 - Not currently, 1 - < 1 pack per day, 2 - $\geq$ 1 pack per day
psmoke	0 - never smoked, 1 - former < 1 pack per day, 2 - former $\geq$ 1 pack per day
pkyrs	Total lifetime packs of cigarettes smoked
crc	0 - No colorectal Cancer, 1 - Colorectal cancer
cayrs	Years to colorectal cancer, or death, or end of follow-up.

Loading Data

We can load this data into R.

```
library(haven)
phscrc <- read_dta("phscrc.dta")
phscrc$age.cat <- cut(phscrc$age, c(40,50,60,70, 90), right=FALSE)
phscrc$alcohol.use <- factor(phscrc$alcohol>0, labels=c("no", "yes"))
```

The Data

```
 mytable <- xtabs(~age.cat+alcohol.use+ crc, phscrc)
library(pander)
pandoc.table(mytable)
```

AGE	ALCOHOL USE	COLORECTAL CANCER	
		No	Yes
40-49	No	1762	3
	Yes	4764	36
50-59	No	1385	19
	Yes	3969	61
60-69	No	749	21
	Yes	2167	73
70-84	No	278	9
	Yes	690	32

Confounding

- The above chart explores the relationship between Alcohol Use, age and Colorectal Cancer.
- We may be interested in evaluating the exposure of Alcohol use on the outcome of Colorectal Cancer adjusting for age.
- We could run the following model:

$$CRC = \beta_0 + \beta_1 E + \beta_2 X$$

- Where $E = 1$ is the group with people who drink more than monthly and X is age in years.

Interpretation of β_1

- The odds ratio between disease and exposure for a fixed age is

$$\frac{\frac{p_{1,x}}{1-p_{1,x}}}{\frac{p_{0,x}}{1-p_{0,x}}}$$

- Thus the log odds ratio is

$$\begin{aligned} & log\left(\frac{p_{1,x}}{1-p_{1,x}}\right) - \log\left(\frac{p_{0,x}}{1-p_{0,x}}\right) \\ &= (\beta_0 + \beta_1(1) + \beta_2 x) - (\beta_0 + \beta_1(0) + \beta_2 x) \\ &= \beta_1 \end{aligned}$$

What does this mean?

- This means
  - For any logistic regression model with a main effect of Exposure, E, and confounder ,X, for any given value of $X = x$, the log odds ratio between disease and exposure is β_1
  - Including both X and E in the model, controls for the possible confounding of X on the relationship between the disease and exposure.
  - If the variable X were omitted, the estimated relationship between disease and exposure could possibly be biased because of confounding by X.

Interpretation of β_2

- For two subjects who have the same exposure E, regardless of which it is, but who differ by 1 year of age, the log(OR) of Colorectal Cancer for a unit change in X is:

$$
\begin{aligned}
&log\left(\frac{p_{e,x+1}}{1-p_{e,x+1}}\right) - \log\left(\frac{p_{e,x}}{1-p_{e,x}}\right) \\
&\quad = (\beta_0 + \beta_1 e + \beta_2(x+1)) - (\beta_0 + \beta_1 e + \beta_2 x) \\
&\quad = \beta_2
\end{aligned}
$$

Effect Modification / effect modification

- Consider the following logistic regression model

$$\log\left(\frac{p_{e,x}}{1-p_{e,x}}\right) = \beta_0 + \beta_1 e + \beta_2 x + \beta_3 \cdot e \cdot x$$

- This gives us a similar model to previously with an added effect modification term.
- This more complex model allows for the odds ratio between outcome and exposure vary across levels of X.

What Happens with Effect Modification?

- For Unexposed (Drink Less than Once a Month):

$$\begin{aligned}\log\left(\frac{p_{0,x}}{1-p_{0,x}}\right) &= \beta_0 + \beta_1 \cdot 0 + \beta_2 x + \beta_3 \cdot 0 \cdot x \\ &= \beta_0 + \beta_2 x\end{aligned}$$

- For Exposed (Drink More than Once a Month):

$$\begin{aligned}\log\left(\frac{p_{1,x}}{1-p_{1,x}}\right) &= \beta_0 + \beta_1 \cdot 1 + \beta_2 x + \beta_3 \cdot 1 \cdot x \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x\end{aligned}$$

What does this mean?

- This implies that the log odds ratio measuring the association between outcome and exposure given $X = x$ is

$$log\left(\frac{p_{1,x}}{1-p_{1,x}}\right) - \log\left(\frac{p_{0,x}}{1-p_{0,x}}\right) = \beta_1 + \beta_3 x$$

- Notice how this varies with different ages now.

Interpretation of the Main effects

- When we have effect modification terms in the model we refer to the E and X as the main effects and $E \cdot X$ as the effect modification term.
  - Main effects do not have their usual interpretation when they are in a model with an effect modification term.
  - $\beta_1 = \beta_1 + \beta_3 \cdot 0$, this represents the log odds ratio of having colorectal cancer comparing this who drink often to those who don't drink often in people whose age is 0.
  - β_2 represents the log odds ratio comparing people who differ in age by 1 year but both drink less than once a month.

Centering Covariates

- One way to make the effect modification terms more meaningful is to center continuous covariates:

$$Z = X - \bar{X}$$

- Then we can fit the model:

$$\log\left(\frac{p_{e,z}}{1-p_{e,z}}\right) = \beta_0^* + \beta_1^* e + \beta_2^* z + \beta_3^* \cdot e \cdot z$$

New Interpretation

- This means that now

$$\beta_1^* = \beta_1^* + \beta_3^* \cdot 0$$

- is the relative odds of having colorectal cancer comparing men who drink more than once a month to those who drink less at the mean age.

Is Effect Modification Significant?

- If there is no effect modification between Age and Alcohol Use pressure than we would find that $\beta_3 = 0$ we can test for this

$$H_0 : \beta_3 = 0 \text{ vs } H_1 : \beta_3 \neq 0$$

The models in R

- We can run these models now:

```
mod1 <- glm(crc ~ alcohol.use+ age, phscrc, family=binomial(link="logit"))
mod2 <- glm(crc ~ alcohol.use*age, phscrc, family=binomial(link="logit"))
phscrc$age.cent <- phscrc$age - mean(phscrc$age,na.rm=TRUE)
mod3 <- glm(crc ~ alcohol.use*age.cent, phscrc, family=binomial(link="logit"))
```

Model 1

- If we look at the first model we find that we have:

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	-8.3348992	0.0000000	-9.1304585	-7.5591081
alcohol.useyes	0.3560274	0.0236256	0.0560273	0.6739932
age	0.0697425	0.0000000	0.0575685	0.0819576

- For 2 people the same age a person who drinks more than once a month has a 42.7646629% increase in odds of colorectal cancer than someone who does not drink.

Model 2

- Then if we consider our model with effect modification

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	-8.5800802	0.000000	-10.2022552	-7.0380756
alcohol.useyes	0.6722965	0.460657	-1.0882345	2.4909110
age	0.0737894	0.000000	0.0482466	0.0995000
alcohol.useyes:age	-0.0052401	0.723872	-0.0344218	0.0238241

- For those who drink less than monthly a one year increase in age leads to a 7.6580044% increase in odds of colorectal cancer.

Model 3

- Finally our model with centered age and effect modification

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	-8.5800802	0.000000	-10.2022552	-7.0380756
alcohol.useyes	0.6722965	0.460657	-1.0882345	2.4909110
age	0.0737894	0.000000	0.0482466	0.0995000
alcohol.useyes:age	-0.0052401	0.723872	-0.0344218	0.0238241

- This time we can interpret the coefficient of alcohol use. For men at the average age in the study, 53.1611937, a man who drinks more than once a month will have 48.2496903% increase in odds compared to a man who drinks less than once a month.

Nested Models

- We move from the base explanation of multiple logistic regression to discussing some tools for comparing models. The first models which we can compare are nested models.
- Nested models can best be explained by the list below:

1. $CRC = \beta_0 + \beta_1 Age$
2. $CRC = \beta_0 + \beta_1 Alcohol + \beta_2 BMI$
3. $CRC = \beta_0 + \beta_1 Age + \beta_2 Alcohol + \beta_3 BMI$

Nested Models

- With these models we have that Model 1 and 2 are nested in Model 3.
- However Model 1 is not next in Model 2.
- With these models the maximized value of the likelihood (and log-likelihood) for the larger model will always be at least as large as that for the smaller model.

Nested Model: Our example

- In our example this means

$$L(3) \geq L(1)$$

$$L(3) \geq L(2)$$

- Where L is the maximized likelihood.

Likelihood Ratio Test

- Like the F test in linear regression we can compare nested models with what is called the Likelihood Ratio Test:

$$\begin{aligned} LR &= 2\log\left(\frac{L(\text{Larger Model})}{L(\text{smaller model})}\right) \\ &= 2\log(L(\text{Larger Model}) - 2\log(L(\text{smaller model})) \end{aligned}$$

The Likelihood Ratio Tes

- This test works with groups of coefficients.
- This means that

$$H_0 : \text{Extra Variables in Larger Model} = 0$$

$$\text{vs}$$

$$H_1 : \text{Extra Variables in Larger Model} \neq 0$$

- Under, H_0 the LR statistic has a χ^2 distribution with d = Number of extra variables in larger model.

Global Null Hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Global Null Example

- We can test the global null hypothesis like in linear regression.
- For example if we consider model 1 from our confounding section

NULL.DEVIANCE	DF.NULL	LOGLIK	AIC	BIC	DEVIANCE	DF.RESIDUAL
2609.171	16017	-1240.335	2486.67	2509.715	2480.67	16015

- We can see that R gives us *Null deviance* and *Residual deviance*. Where
  - *Null Deviance* = 2(LL(Saturated Model) - LL(Null Model)) on $df = df_{Sat} - df_{Null}$
  - *Residual Deviance* = 2(LL(Saturated Model) - LL(Proposed Model)) on $df = df_{Sat} - df_{Proposed}$

Saturated Model

- The Saturated model means that each data point has its own probability.
- The null model is a model with just an intercept term or essentially the probability of getting a disease is the same for everyone.
- The proposed model is the model that we are fitting.

Likelihood Ratio Test

- We can then do a likelihood ratio test in R to test whether $\beta_1 = \beta_2 = 0$.
- We do this by comparing

$$\text{Null Deviance} - \text{Residual Deviance} \sim \chi^2_d$$

- Where $d = p$ or the number of covariates in the model.

Manually in R

```
LR <- summary(mod1)$null.deviance - summary(mod1)$deviance
df <-   summary(mod1)$df.null - summary(mod1)$df.residual
 1- pchisq(LR,df)
```

```
## [1] 0
```

- We find that this gives us a p-value < 0.0001 so that at least β_1 or β_2 is significant.

Categorical Covariates

Many times we have categorical covariates and we need to be able to use them in model building. We can proceed two different ways depending on the data.

If we have covariate X which is categorical than we treat X differently depending on

1. If the categories of X are **nominal** or unordered (race, gender, eye color, …)
2. If the categories of X can be ordered.

Nominal Categories.

With X which has $K \geq 2$ categories, we created $K - 1$ variables. For example:

- $x_1 = 1$ if $X = 1$, $x_1 = 0$ otherwise
- $x_2 = 1$ if $X = 2$, $x_2 = 0$ otherwise
- $x_{k-1} = 1$ if $X = 2$, $x_{k-1} = 0$ otherwise

Nominal Categories

- This means that if a person was in category K then $x_1, x_2, \ldots, x_{k-1} = 0$.
- We would then fit a logistic model:

$$logit(\Pr(Y = 1|X)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1}$$

Indicator Variables

- Many times we write this with *Indicator Variables* which is
  - I(Cat=1) is 1 if in category 1 and 0 otherwise
  - I(Cat=2) is 1 if in category 2 and 0 otherwise
  - I(Cat=K-1) is 1 if in category K-1 and 0 otherwise

What is the Saturated Model?

- This would be the same model as above.
- With this model we have a saturated model because our model contains K predictors which is the number of levels of X.
- This would be the same model as for a $2 \times K$ table in epidemiology.

We could then have:

CATEGORY	LOG ODDS	PROBABILITY
K	β_0	$p_0 = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$
1	$\beta_0 + \beta_1$	$p_0 = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$
2	$\beta_0 + \beta_2$	$p_0 = \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$
$\vdots$	$\vdots$	$\vdots$
K-1	$\beta_0 + \beta_{k-1}$	$p_0 = \frac{\exp(\beta_0+\beta_{k-1})}{1+\exp(\beta_0+\beta_{k-1})}$

Ordered Categories

- For ordered categories we can treat them like before or we can treat them as a trend.
- In order to do so let us consider what a trend looks like.
- If we have a trend what we are saying is that if you consider the probabilities of disease in the K categories than

$$p_1 \geq p_2 \geq \cdots \geq P_K \text{ or } p_1 \leq p_2 \leq \cdots \leq P_K$$

Trend Variable

- Let us consider the original variable of `Alcohol` and its effect of Colorectal Cancer.

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	-4.4182131	0.0000000	-4.6563690	-4.1931173
alcohol	0.2776848	0.0019606	0.1023961	0.4541972

Alcohol Categories

$$\text{Alcohol} = \begin{cases} 0 & \text{Less than monthly} \\ 1 & \text{Monthly to less than Daily} \\ 2 & \text{Daily or more} \end{cases}$$

What Does the Trend mean?

- If we consider the above model what we have is

$$logit(p_x) = \beta_0 + \beta_1 * Alcohol$$

- We then have

$$\begin{aligned} logit(p_0) &= \beta_0 \\ logit(p_1) &= \beta_0 + 1\beta_1 \\ logit(p_2) &= \beta_0 + 2\beta_2 \end{aligned}$$

What does this mean?

This means if

$$\beta_1 = 0 \Rightarrow p_0 = p_1 = p_2$$
$$\beta_1 > 0 \Rightarrow p_0 < p_1 < p_2$$
$$\beta_1 < 0 \Rightarrow p_0 > p_1 > p_2$$

What are we testing?

- So a test for $\beta_1 = 0$ will show us if the natural order of our categories is correct.
- We can also test whether we should use the Linear Trend vs the Indicator Variable Model.
- To do this we consider the Likelihood Ratio Test.
- with the following models:
  - $logit(p_x) = \beta_0 + \beta_1 Alcohol$
  - $logit(p_x) = \beta_0 + \beta_1 I(Alcohol = 1) + \beta_2 I(Alcohol = 2)$

How do the Models compare?

-We will find that Model 1 is nested within Model 2. - How does this work? - $\beta_1 = \beta_2$ model 2.

Results of Factor Model

TERM	ESTIMATE	P.VALUE	CONF.LOW	CONF.HIGH
(Intercept)	-4.3853864	0.0000000	-4.6716150	-4.1235215
factor(alcohol)1	0.2173767	0.1928240	-0.1035724	0.5522374
factor(alcohol)2	0.5436799	0.0024019	0.1961081	0.8999708

Likelihood Ratio Test

```
anova_mod <-anova(mod4, mod5, test="Chisq")
knitr::kable(tidy(anova_mod))
```

RESID..DF	RESID..DEV	DF	DEVIANCE	P.VALUE
16016	2599.51	NA	NA	NA
16015	2599.33	1	0.1807433	0.6707353

- I find that with a p-value of 0.6707353.
- I fail to reject the null hypothesis and find that I can stick with the linear trend rather than the indicator variable model.