BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 4 supplementary slides for `R` enthusiasts

Version: 01/25/2021

Examples for Set 4

Examples for R enthusiasts:

- Diabetes and gender in MRI data (Slide 395)
- Diabetes and race in MRI data (Slide 432)
- Diabetes and CHD in MRI data (Slide 471)
- General health in MRI data (Slide 477)
- Age and lymph nodes in endometrial study (Slide 486)

Examples for Set 4

Examples for R enthusiasts:

- **Diabetes and gender in MRI data (Slide 395)**
- Diabetes and race in MRI data (Slide 432)
- Diabetes and CHD in MRI data (Slide 471)
- General health in MRI data (Slide 477)
- Age and lymph nodes in endometrial study (Slide 486)

Diabetes and gender in MRI study

Example:

- X: 0 = female; 1 = male.
- Y: 0 = no diabetes; 1 = diabetes.

	Diabetes	No diabetes	Total
Male	53	313	366
Female	26	343	369
Total	79	656	735

- Estimated prevalence difference: 0.0743
- Estimated odds ratio (OR): 2.234
- Estimated prevalence ratio (RR): 2.055

Diabetes and gender in MRI study

Reading in the MRI data:

- Read in data:

```
mri.data <- read.csv("mri.csv",
                     header = TRUE,
                     stringsAsFactors = FALSE)
```

Diabetes and gender in MRI study

Binary outcome regression: Identity link

- Function `glm` in R (must specify family and link):

```
model.1 <- glm(diabetes ~ male,
               family = binomial(link = "identity"),
               data = mri.data)
```

Diabetes and gender in MRI study

Binary outcome regression: Identity link (Results)

```
> summary(model.1)

Call:
glm(formula = diabetes ~ male, family = binomial(link = "identity"),
    data = mri.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5593  -0.5593  -0.3823  -0.3823   2.3034

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.07046    0.01332   5.289 1.23e-07 ***
male         0.07435    0.02271   3.273  0.00106 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 501.59  on 734  degrees of freedom
Residual deviance: 490.82  on 733  degrees of freedom
AIC: 494.82

Number of Fisher Scoring iterations: 2
```

- Coefficient estimates agree with Stata output (Slide 396).

Diabetes and gender in MRI study

Binary outcome regression: Identity link

- Sandwich variance:

```
robust.var <- vcovHC(model.1, type = "HC1")

## Output
> sqrt(diag(robust.var))
(Intercept)        male
 0.01334092  0.02274339
```

- Does *not* agree perfectly with Stata output (Slide 396).
- The reason is a different degrees of freedom correction.
  - Stata uses $N - 1$; R uses $N - K$ ($K = 2$ in this case).

Diabetes and gender in MRI study

Binary outcome regression: Identity link

- Sandwich variance (calibrating degrees of freedom):

```
N <- dim(model.1$model)[1]
robust.var.df <- vcovHC(model.1, type = "HC1") * ((N - 2)/(N - 1))

## Output
> sqrt(diag(robust.var.df))
(Intercept)        male
 0.01333183  0.02272789
```

- Agrees with Stata output (Slide 396).

Diabetes and gender in MRI study

Side note:

- Stata seems to often use $N - 1$ irrespective of the number of model parameters.
- If you are using R for assignments, you are not expected to change the degrees of freedom correction to match Stata's. I'm just illustrating why there is a discrepancy here.

Diabetes and gender in MRI study

Binary outcome regression: Logit link

- Function `glm` in R (must specify family and link):

```
model.2 <- glm(diabetes ~ male,
               family = binomial(link = "logit"),
               data = mri.data)
robust.var <- vcovHC(model.2, type = "HC1") * (N - 2)/(N - 1)
```

Diabetes and gender in MRI study

Binary outcome regression: Logit link (salient results)

```
exp(c(OR = coef(model.2)[2],
      CI.Low = coef(model.2)[2] - qnorm(0.975) * sqrt(diag(robust.var))[2],
      CI.High = coef(model.2)[2] + qnorm(0.975) * sqrt(diag(robust.var))[2]))

## Output
    OR.male  CI.Low.male CI.High.male
    2.233841     1.363051     3.660940
```

- Agrees with Stata output (Slide 397).

Diabetes and gender in MRI study

Binary outcome regression: Log link

- Function `glm` in R (must specify family and link):

```
model.3 <- glm(diabetes ~ male,
               family = binomial(link = "log"),
               data = mri.data)
robust.var <- vcovHC(model.3, type = "HC1") * ((N - 2)/(N - 1))
```

Diabetes and gender in MRI study

Binary outcome regression: Log link (salient results)

```
exp(c(RR = coef(model.3)[2],
      CI.Low = coef(model.3)[2] - qnorm(0.975) * sqrt(diag(robust.var))[2],
      CI.High = coef(model.3)[2] + qnorm(0.975) * sqrt(diag(robust.var))[2]))

## Output
    RR.male  CI.Low.male CI.High.male
    2.055170     1.314688     3.212721
```

- Agrees with Stata output (Slide 398).

Examples for Set 4

Examples for R enthusiasts:

- *Diabetes and gender in MRI data (Slide 395)*
- **Diabetes and race in MRI data (Slide 432)**
- Diabetes and CHD in MRI data (Slide 471)
- General health in MRI data (Slide 477)
- Age and lymph nodes in endometrial study (Slide 486)

Diabetes and race in MRI study

Example:

- X: 1 = white; 2 = black; 3 = Asian; 4 = other.
- Y: 0 = no diabetes; 1 = diabetes.
- Model:

$$\log(\mathrm{P}(Y = 1|X = x)) = \beta_0 + \beta_1 1(x = 2) + \beta_2 1(x = 3) + \beta_3 1(x = 4)$$

- Hypothesis test: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_1 :$ (not H_0).

Diabetes and race in MRI study

Binary outcome regression: Joint testing

- Fit model and extract robust variance (note that we're re-calibrating the degrees of freedom correction).

```
model.race <- glm(diabetes ~ factor(race),
                  family = binomial(link = "log"),
                  data = mri.data)

N <- dim(model.race$model)[1]
robust.var <- vcovHC(model.race, type = "HC1") * ((N - 4)/(N - 1))
```

Diabetes and race in MRI study

Binary outcome regression: Joint testing

- The `testparm.R` function will work in this context.

```
testparm.R(par = list(2,3,4),
           coefs = coef(model.race),
           vcov = robust.var,
           type = "W")

## Output
         W           P
6.62942677 0.08469562
```

- Agrees with Stata output (Slide 433).
  - Note the use of a Wald test rather than an F-test.

Examples for Set 4

Examples for R enthusiasts:

- *Diabetes and gender in MRI data (Slide 395)*
- *Diabetes and race in MRI data (Slide 432)*
- **Diabetes and CHD in MRI data (Slide 471)**
- General health in MRI data (Slide 477)
- Age and lymph nodes in endometrial study (Slide 486)

Diabetes and CHD in MRI study

Example:

- Multinomial regression model using MRI data:
  - X_1: 0 = no diabetes; 1 = diabetes.
  - X_2: age (years).
  - X_3: 0 = female; 1 = male.
  - Y: 0 = no CHD; 1 = angina; 2 = myocardial infarction.

Diabetes and CHD in MRI study

Multinomial regression:

- The `multinom` function is the most reliable one I could find, and requires the `nnet` package.

```
mreg <- multinom(chd ~ diabetes + age + male,
                 data = mri.data)
```

Diabetes and CHD in MRI study

Multinomial regression: Results

```
> exp(summary(mreg)$coefficients)

  (Intercept) diabetes      age     male
1  0.00250757 1.095997 1.048926 1.431986
2  0.06102624 1.773780 1.006461 2.003295
```

- Estimated RRRs agree with Stata output (Slide 472).

Diabetes and CHD in MRI study

Multinomial regression: Testing

- I am unaware of a method to obtain robust standard errors with `multinom` other than hard-coding. In this class, not worth effort to hard-code robust standard error for this model.
- Note: `testparm.R` works with non-robust variance (`vcov`).

```
testparm.R(par = list(2,6),
           coefs = c(coef(mreg)[1,],
                     coef(mreg)[2,]),
           vcov = vcov(mreg),
           type = "W")

## Output
         W          P
3.3446370 0.1878111
```

- Does not agree exactly with Stata output (Slide 474).

Examples for Set 4

Examples for R enthusiasts:

- *Diabetes and gender in MRI data (Slide 395)*
- *Diabetes and race in MRI data (Slide 432)*
- *Diabetes and CHD in MRI data (Slide 471)*
- **General health in MRI data (Slide 477)**
- Age and lymph nodes in endometrial study (Slide 486)

General health in MRI study

Example:

- Proportional odds model:
  - X_1: age (years).
  - X_2: 0: female; 1: male.
  - Y: view of own health (1:5)
    - Higher values indicate poorer view of health.

General health in MRI study

Ordinal regression:

- The `polr` function is the most reliable one I could find, and requires the `MASS` package.

```
model.gh <- polr(factor(genhlth) ~ age + male,
                 data = mri.data)
```

General health in MRI study

Ordinal regression: Results

```
exp(summary(model.gh)$coef[1:2,1])
```

```
## Output
      age      male
1.0277373 0.9203347
```

- Odds ratios do not agree perfectly with Stata output (Slide 478), but they are close.
  - Reason for discrepancy not clear (likely numeric in nature rather than the result of a substantive modeling assumption).

General health in MRI study

Ordinal regression: Standard errors

- The `vcovHC` function is not compatible with `polr` command, but the `sandwich` function is (does not include a degrees of freedom correction).

```
N <- dim(model.gh$model)[1]

> sqrt(diag(sandwich(model.gh) * (N)/(N - 1)))[1:2]

Re-fitting to get Hessian

       age       male
0.01339874 0.13542919
```

- Does not agree perfectly with Stata output (Slide 478), but they are close.

Examples for Set 4

Examples for R enthusiasts:

- *Diabetes and gender in MRI data (Slide 395)*
- *Diabetes and race in MRI data (Slide 432)*
- *Diabetes and CHD in MRI data (Slide 471)*
- *General health in MRI data (Slide 477)*
- **Age and lymph nodes in endometrial study (Slide 486)**

Age and lymph nodes in endometrial study

Example:

- Y: # of nodes removed (count).
- X: age (years).
- Model: $\log(\mathrm{E}[Y|X = x]) = \beta_0 + \beta_1 x$

Age and lymph nodes in endometrial study

Reading in the endometrial data:

- Read in data:

```
endo.data <- read.csv("endometrial.csv",
                      header = TRUE,
                      stringsAsFactors = FALSE)
```

Age and lymph nodes in endometrial study

Poisson regression: Log link

- Function `glm` in R (must specify family and link):

```
model.nodes <- glm(nodes ~ age,
                   family = poisson(link = "log"),
                   data = endo.data)
```

Age and lymph nodes in endometrial study

Poisson regression: Log link

- Results:

```
N <- dim(regr.pois$model)[1]
robust.var <- vcovHC(regr.pois, type = "HC1") * ((N - 2)/(N - 1))

exp(c(IRR = coef(model.nodes)[2],
    CI.Low = coef(model.nodes)[2] - qnorm(0.975) * sqrt(diag(robust.var))[2],
    CI.High = coef(model.nodes)[2] + qnorm(0.975) * sqrt(diag(robust.var))[2]))

## Output
    IRR.age  CI.Low.age CI.High.age
   1.012862    1.002544    1.023286
```

- Agrees with Stata output (Slide 488).