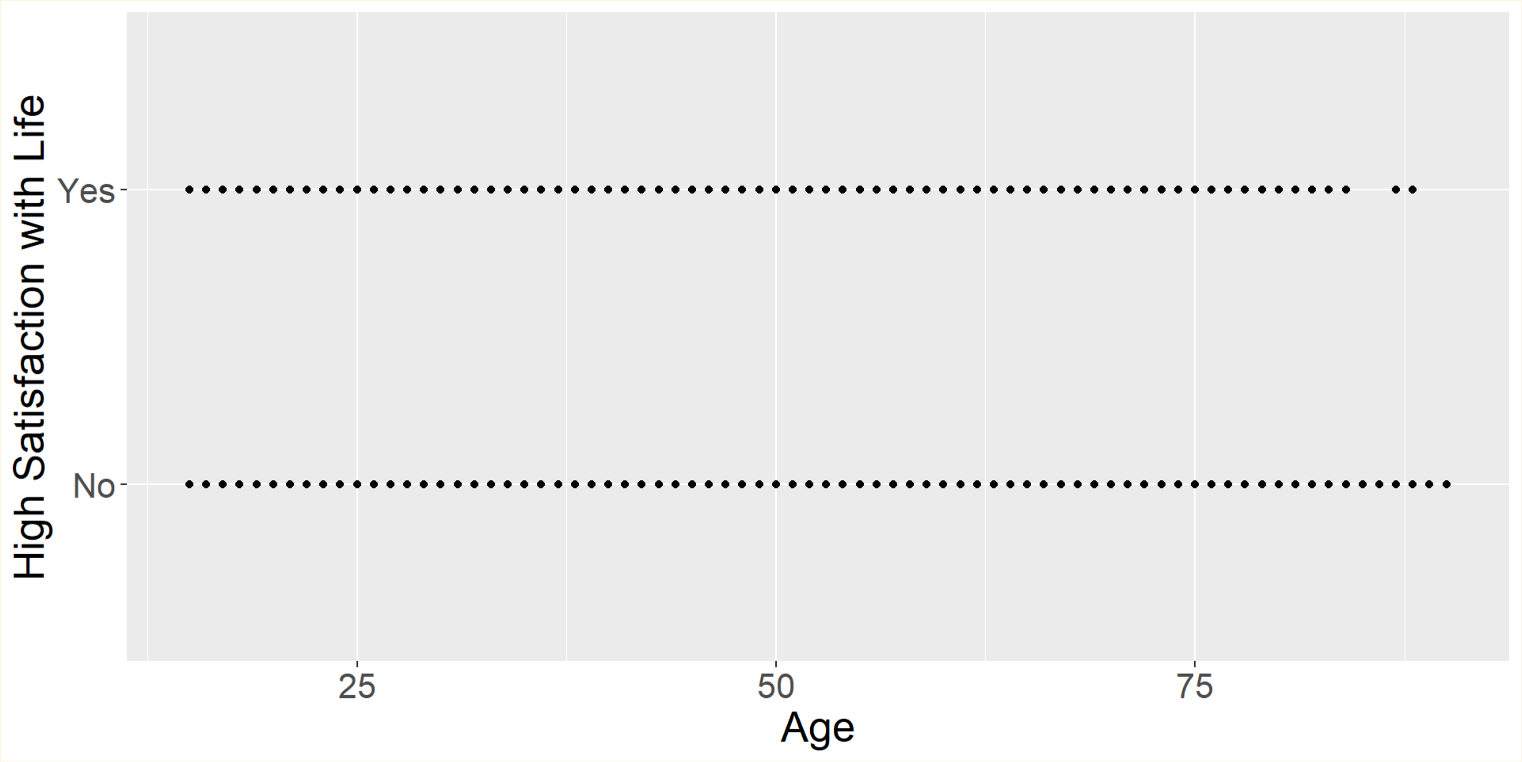
GENERALIZED LINEAR MODELS FOR BINARY SURVEY VARIABLES
LOGISTIC REGRESSION

Yongchao Ma
ytma@umich.edu

Jul 1 2024

WHY LOGISTIC REGRESSION

- When the dependent variable can only be 0 or 1, do we want to draw a straight line through?



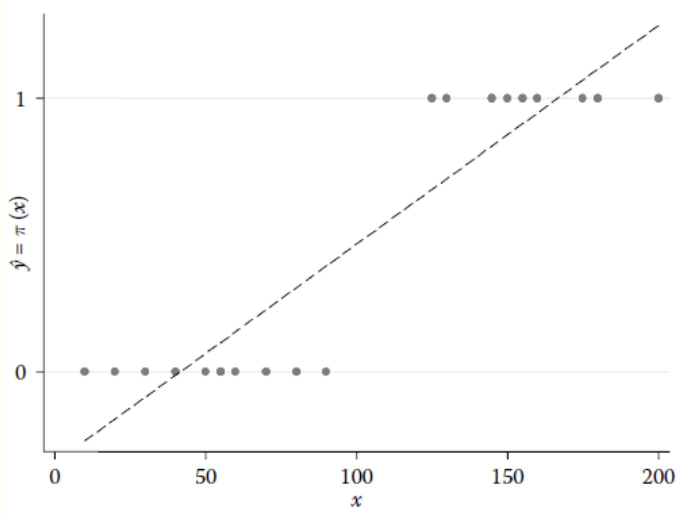
- Rather than predicting these two values, we try to model the *probabilities* that the dependent variable takes one of these two values

LINEAR TO LOGISTIC REGRESSION

- Consider a linear regression

$$\boldsymbol{\pi} = \Pr(y = 1|\mathbf{x}) = B_0 + B_1 x_1 + \cdots + B_p x_p + e$$

- The probability must be between 0 and 1, but the linear predictor $\boldsymbol{\eta} = B_0 + B_1 x_1 + \cdots + B_p x_p$ on the right hand side can take any real number



- Transform the probability to remove the range restrictions, and model the transformation as a linear function of the covariates

LINEAR TO LOGISTIC REGRESSION

- First, we move from the probability $[0, 1]$ to the odds $(0, \infty)$

$$\text{odds} = \frac{\Pr(y = 1|\mathbf{x})}{\Pr(y = 0|\mathbf{x})} = \frac{\boldsymbol{\pi}}{1 - \boldsymbol{\pi}} = e^{B_0 + B_1 x_1 + \cdots + B_p x_p}$$

- Second, we take logarithms to move from the odds to the log-odds $(-\infty, +\infty)$

$$\text{logit}(\boldsymbol{\pi}) = \ln\left(\frac{\boldsymbol{\pi}}{1 - \boldsymbol{\pi}}\right) = B_0 + B_1 x_1 + \cdots + B_p x_p$$

- Logarithmic transformation maps probabilities from the range $[0, 1]$ to the entire real line. The probability can be solved from the logit model

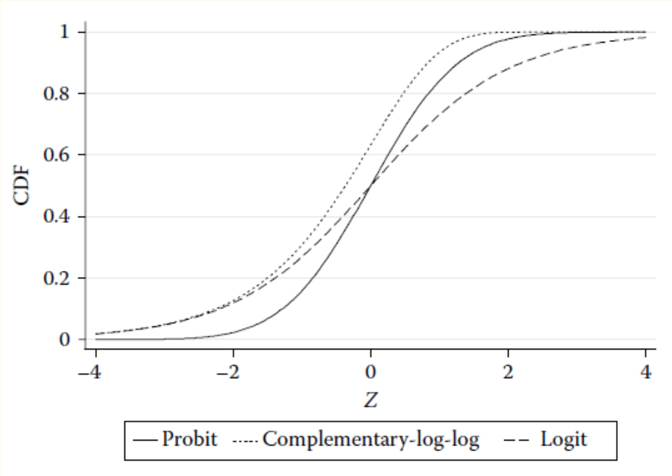
$$\boldsymbol{\pi} = \frac{e^{B_0 + B_1 x_1 + \cdots + B_p x_p}}{1 + e^{B_0 + B_1 x_1 + \cdots + B_p x_p}} = \frac{1}{1 + e^{-(B_0 + B_1 x_1 + \cdots + B_p x_p)}}$$

LINEAR TO LOGISTIC REGRESSION

- The binary dependent variable y is assumed to follow a binomial distribution
  - $\mathrm{E}(y) = n\pi$
  - $\mathrm{Var}(y) = n\pi(1 - \pi)$
  - The mean and variance depend on the underlying probability π
  - Any covariate x that affects the probability also affects both the mean and variance
- Normality and homoscedasticity assumptions are violated
  - Least squares estimation is not appropriate
  - Maximum likelihood estimation is used

GENERALIZED LINEAR MODELS

- Generalized linear models have three components
  - Distribution of the dependent variable
  - Linear predictor $\boldsymbol{\eta} = B_0 + B_1 x_1 + \cdots + B_p x_p$
  - Link function: the transformation that describes how the mean of the dependent variable is related to the linear predictor $g\left(\mathrm{E}(y|\mathbf{x})\right) = \boldsymbol{\eta}$
    - Logit $g(\boldsymbol{\pi}) = \ln\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right)$
    - Probit $g(\boldsymbol{\pi}) = \Phi^{-1}(\boldsymbol{\pi})$ where Φ is the standard normal cumulative distribution function
    - Complementary log-log $g(\boldsymbol{\pi}) = \ln(-\ln(1-\boldsymbol{\pi}))$



MODEL ESTIMATION UNDER SRS

- The logistic regression model is estimated using *maximum likelihood estimation*
- The likelihood function for a SRS of n independent binomial observations is the product of the probabilities of observing the data given the parameters

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} (\pi_i)^{y_i} (1 - \pi_i)^{1-y_i}$$

- Estimate parameters
  - Take the 1st derivative of $\ln L(\boldsymbol{\beta})$ with respect to each parameter, set them to zero, and solve for the parameters
  - No closed-form solution, iterative methods (e.g., Newton-Raphson algorithm) are used
- Estimate variance of parameter estimates
  - Take the 2nd derivative of $\ln L(\boldsymbol{\beta})$ with respect to each parameter, evaluate at the maximum likelihood estimate, and invert to obtain the variance-covariance matrix of the parameter estimates

MODEL ESTIMATION UNDER COMPLEX SAMPLING

- Maximum likelihood estimation is not appropriate for complex sample designs
  - probability of selection is not constant across observations
  - observations are not independent due to clustering or stratification

- Consider complex sampling from a finite population, *pseudo-maximum likelihood estimation* is used to estimate regression parameters

$$PL(\mathbf{B}) = \prod_{i=1}^{n} \left((\pi_i)^{y_i} (1 - \pi_i)^{1-y_i} \right)^{w_i}$$

with

$$\pi_i = \frac{e^{x_i \mathbf{B}}}{1 + e^{x_i \mathbf{B}}}$$

MODEL ESTIMATION UNDER COMPLEX SAMPLING

- Estimate parameters
  - Maximize the weighted pseudo-likelihood function using the iterative method as in the standard maximum likelihood estimation
- Estimate variance of parameter estimates
  - Taylor series estimation
  - Replication methods (JRR or BRR)

TESTS OF MODEL PARAMETERS

- Wald test
  - Test the null hypothesis that a parameter is equal to 0
  - The test statistic is the ratio of the parameter estimate to its standard error
  - The test statistic is referred to Student t distribution with design-based degrees of freedom
  - Alternatively, we can treat the square of the test statistic as a χ^2 statistic with one degree of freedom

- Likelihood ratio test
  - Compare the likelihood of the model with the parameter of interest to the likelihood of the model without the parameter of interest
  - The test statistic is twice the difference in the log-likelihoods of the two models
  - The test statistic is referred to a χ^2 distribution with the difference in the number of parameters between the two models
  - *Not applicable* to complex sample designs

REGRESSION DIAGNOSTICS

- Goodness of Fit (for SRS)
  - Pearson, Deviance
  - Hosmer-Lemeshow test
  - Classification table
  - Area under the ROC curve
  - Psuedo-R^2
- Influence and Outliers
  - Cook's distance
  - "Hat" matrix
  - Change in χ^2 statistic due to deletion of observations

EXAMPLE: LIFETIME MAJOR DEPRESSION

- Data: National Comorbidity Survey Replication (NCS-R)
- Question: Assess the significance of potential predictors of having lifetime major depression for adults greater than 17 years of age
- Dependent variable:
  - Lifetime major depression (1=Yes; 0=No)
- Predictors:
  - Age (1=18–29; 2=30–44; 3=45–59; 4=60+)
  - Sex (1=Male; 2=Female)
  - Alcohol dependence (1=Yes; 0=No)
  - Education (1=0–11; 2=12; 3=13–15; 4=16+ years)
  - Marital status (1=Married; 2=Previously Married; 3=Never Married)

EXAMPLE: LIFETIME MAJOR DEPRESSION

BIVARIATE ANALYSIS

```r
# Specify survey design
ncsrsvyp2 <- svydesign(id = ~seclustr, strata = ~sestrat,
                       weights = ~ncsrwtlg, data = ncsrp2, nest = TRUE)

# Bivariate chisq tests
svyby(~mdec, ~sexc, design = ncsrsvyp2, svymean)
```

```
         sexc    mdecNo   mdecYes    se.mdecNo   se.mdecYes
Male     Male 0.8471074 0.1528926 0.009137590 0.009137590
Female Female 0.7738295 0.2261705 0.006727609 0.006727609
```

```r
svychisq(~mdec + sexc, design = ncsrsvyp2)
```

```

	Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~mdec + sexc, design = ncsrsvyp2)
F = 44.834, ndf = 1, ddf = 42, p-value = 3.965e-08
```

EXAMPLE: LIFETIME MAJOR DEPRESSION

ODDS RATIO

	% Having lifetime major depression	Odds
Male	0.1528926	$\frac{0.1528926}{1-0.1528926} = 0.1804879$
Female	0.2261705	$\frac{0.2261705}{1-0.2261705} = 0.2922743$

- Odds ratio

$$\mathrm{OR} = \frac{\mathrm{odds}_{\mathrm{female}}}{\mathrm{odds}_{\mathrm{male}}} = \frac{0.2922743}{0.1804879} = 1.619$$

- The odds of having lifetime major depression are 1.619 times higher for females than for males

EXAMPLE: LIFETIME MAJOR DEPRESSION

MODEL SPECIFICATION AND ESTIMATION: SINGLE PREDICTOR

```r
model1 <- svyglm(mdec ~ sexc,
                 design = ncsrsvyp2, family = quasibinomial)
summary(model1)
```

```
Call:
svyglm(formula = mdec ~ sexc, design = ncsrsvyp2, family = quasibinomial)

Survey design:
svydesign(id = ~seclustr, strata = ~sestrat, weights = ~ncsrwtlg,
    data = ncsrp2, nest = TRUE)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.71209    0.07055  -24.27  < 2e-16 ***
sexcFemale   0.48203    0.07237    6.66 4.98e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.000176)
```

EXAMPLE: LIFETIME MAJOR DEPRESSION

INTERPRETATION: SINGLE PREDICTOR

- Logistic regression coefficients represent the change in the log-odds of the dependent variable for a one-unit increase in the predictor, holding all other variables constant
  - The coefficient for *Sex* is 0.48203
  - This means that being female (compared to the reference category, which is male) is associated with an increase in the log-odds of having lifetime major depression

- To make the interpretation more intuitive, we often exponentiate the coefficient to obtain the odds ratio

$$\text{OR} = \frac{\text{odds}_{\text{female}}}{\text{odds}_{\text{male}}} = \frac{e^{B_0+B_1}}{e^{B_0}} = e^{B_1}$$

EXAMPLE: LIFETIME MAJOR DEPRESSION

INTERPRETATION: SINGLE PREDICTOR

```
exp(model1$coef)
```

```
(Intercept)   sexcFemale
  0.1804879    1.6193566
```

- The odds ratio indicates that the odds of having lifetime major depression are 1.619 times higher for females than for males

EXAMPLE: LIFETIME MAJOR DEPRESSION

MODEL SPECIFICATION AND ESTIMATION: MULTIPLE PREDICTORS

```r
model2 <- svyglm(mdec ~ sexc + ag4catc + aldc + ed4catc + mar3catc,
                      design = ncsrsvyp2, family = quasibinomial)
summary(model2)
```

```
Call:
svyglm(formula = mdec ~ sexc + ag4catc + aldc + ed4catc + mar3catc,
    design = ncsrsvyp2, family = quasibinomial)

Survey design:
svydesign(id = ~seclustr, strata = ~sestrat, weights = ~ncsrwtlg,
    data = ncsrp2, nest = TRUE)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -2.16042    0.15214 -14.200 2.30e-15 ***
sexcFemale                   0.57735    0.07722   7.477 1.64e-08 ***
ag4catc30-44                 0.25562    0.09438   2.708   0.0108 *
ag4catc45-59                 0.20645    0.09153   2.256   0.0311 *
```

EXAMPLE: LIFETIME MAJOR DEPRESSION

INTERPRETATION: MULTIPLE PREDICTORS

```r
exp(model2$coef)
```

```
              (Intercept)                    sexcFemale 
                0.1152765                     1.7813032 
             ag4catc30-44                  ag4catc45-59 
                1.2912600                     1.2293019 
               ag4catc60+                       aldcYes 
                0.5087563                     4.1523575 
                ed4catc12                  ed4catc13-15 
                1.0824803                     1.2592434 
               ed4catc16+ mar3catcPreviously Married 
                1.1769489                     1.6264870 
    mar3catcNever Married 
                1.1225236 
```

- The odds ratio indicates that the odds of having lifetime major depression are 1.781 times higher for females than for males, holding all other variables constant

EXAMPLE: LIFETIME MAJOR DEPRESSION

TESTS OF MODEL PARAMETERS

```
# Wald test
regTermTest(model2, ~ag4catc)
```

```
Wald test for ag4catc
 in svyglm(formula = mdec ~ sexc + ag4catc + aldc + ed4catc + mar3catc,
    design = ncsrsvyp2, family = quasibinomial)
F =  19.98292  on  3  and  32  df: p= 1.7536e-07
```

SOFTWARE PROCEDURES USING STATA

- The `svy: logit` and `svy: logistic` commands
  - `svy: logistic` defaults to odds ratio output; coef option yields logistic model parameter estimates
  - `svy: logit` defaults to log-odds (B) output; or option yields odds ratios "i." prefix defines categorical predictors
  - Default to lowest alphanumeric category for reference
  - Change reference category for categorical predictors using ib#.
  - Post-estimation `test` statement for Wald tests of multi-parameter hypotheses
    - e.g. (for ASDA Chapter 8 example), to test the null hypothesis that all of the parameters associated with Age are equal to zero, use this test statement:
    - `test  2.ag4cat 3.ag4cat 4.ag4cat`

SOFTWARE PROCEDURES USING STATA

BIVARIATE ANALYSIS

```
svyset seclustr [pweight = ncsrwtlg], strata(sestrat)
svy: tab ag4cat mde, row
svy: tab sex mde, row
svy: tab ald mde, row
svy: tab ed4cat mde, row
svy: tab mar3cat mde, row
```

SOFTWARE PROCEDURES USING STATA

MODEL SPECIFICATION AND ESTIMATION

```
svy: logit mde i.ag4cat ib2.sex ald i.ed4cat i.mar3cat

* Estimated odds ratios and 95% CIs can be generated in svy: logit by addin

svy: logit mde i.ag4cat ib2.sex ald i.ed4cat i.mar3cat, or
```

SOFTWARE PROCEDURES USING STATA

WALD TESTS OF MULTI-PARAMETER PREDICTORS

```
test 2.ag4cat 3.ag4cat 4.ag4cat
test 2.mar3cat 3.mar3cat
test 2.ed4cat 3.ed4cat 4.ed4cat
```

SOFTWARE PROCEDURES USING STATA

TEST OVERALL GOODNESS OF FIT

- Use Archer and Lemeshow's (2006, 2007) design-adjusted test to assess the goodness of fit of this initial model
- `estat gof` (post-estimation command)
- The resulting design-adjusted F-statistic reported in Stata is equal to $F_{A-L} = 1.229$, with a p-value of 0.310
  - This suggests that the null hypothesis that the model fits the data well is not rejected
  - We therefore have confidence moving forward that the fit of this initial model is reasonable
- Not presently "canned" in R

SOFTWARE PROCEDURES USING STATA

COMPARE LOGIT, PROBIT, AND COMPLEMENTARY LOG-LOG MODELS

```
svy: logit ald i.ag4cat ib2.sex i.ed4cat i.mar3cat
svy: probit ald i.ag4cat ib2.sex i.ed4cat i.mar3cat
svy: cloglog ald i.ag4cat ib2.sex i.ed4cat i.mar3cat
```

SOFTWARE PROCEDURES USING STATA

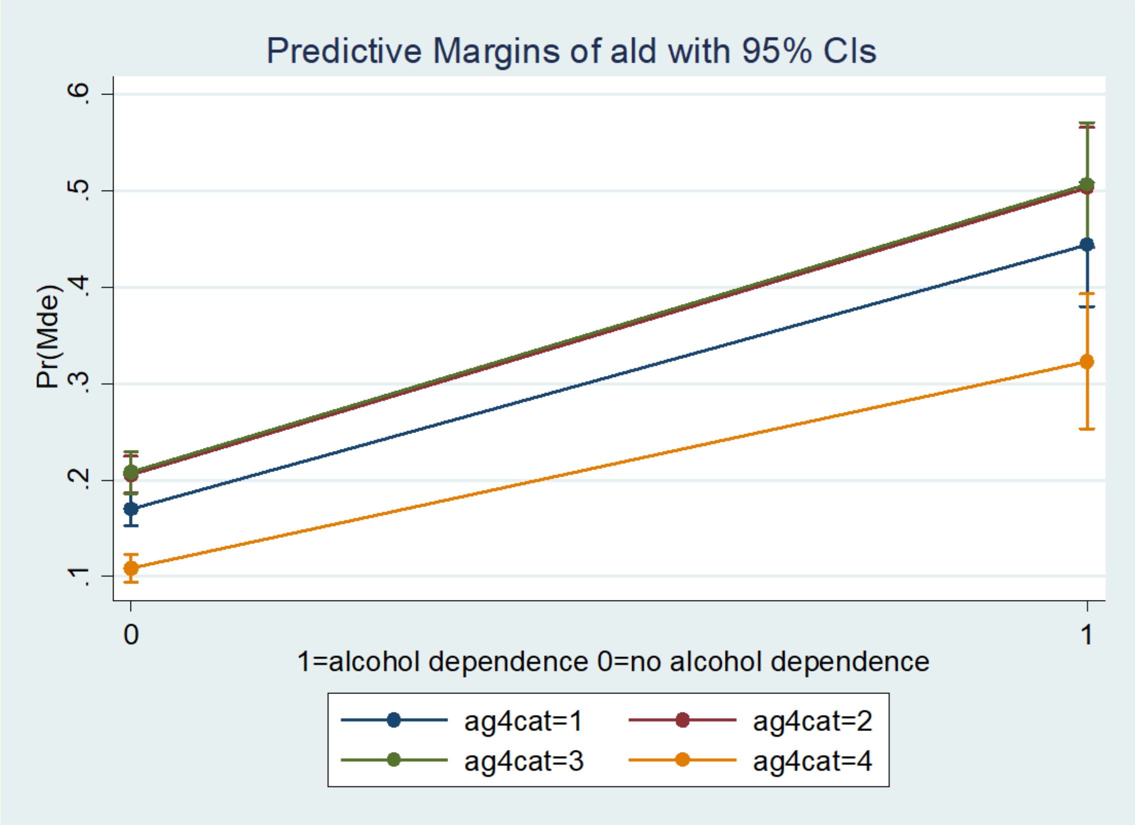
PLOTTING PREDICTED MARGINAL PROBABILITIES AND EFFECTS

- Stata offers extremely easy-to-use post-estimation commands for calculation and plotting of marginal predicted probabilities based on fitted models (not straightforward in R!)
- Default calculation: compute a model-based predicted probability for everyone in the data set as if they all belonged to the same subgroup, and average the predictions
- One can plot marginal predicted probabilities for different subgroups, or average marginal effects (i.e., expected changes in predicted probabilities associated with a one-unit increase in a given predictor)
- The next few slides present some of the examples illustrated in Chapter 8 of ASDA

SOFTWARE PROCEDURES USING STATA

PLOTTING PREDICTED MARGINAL PROBABILITIES AND EFFECTS

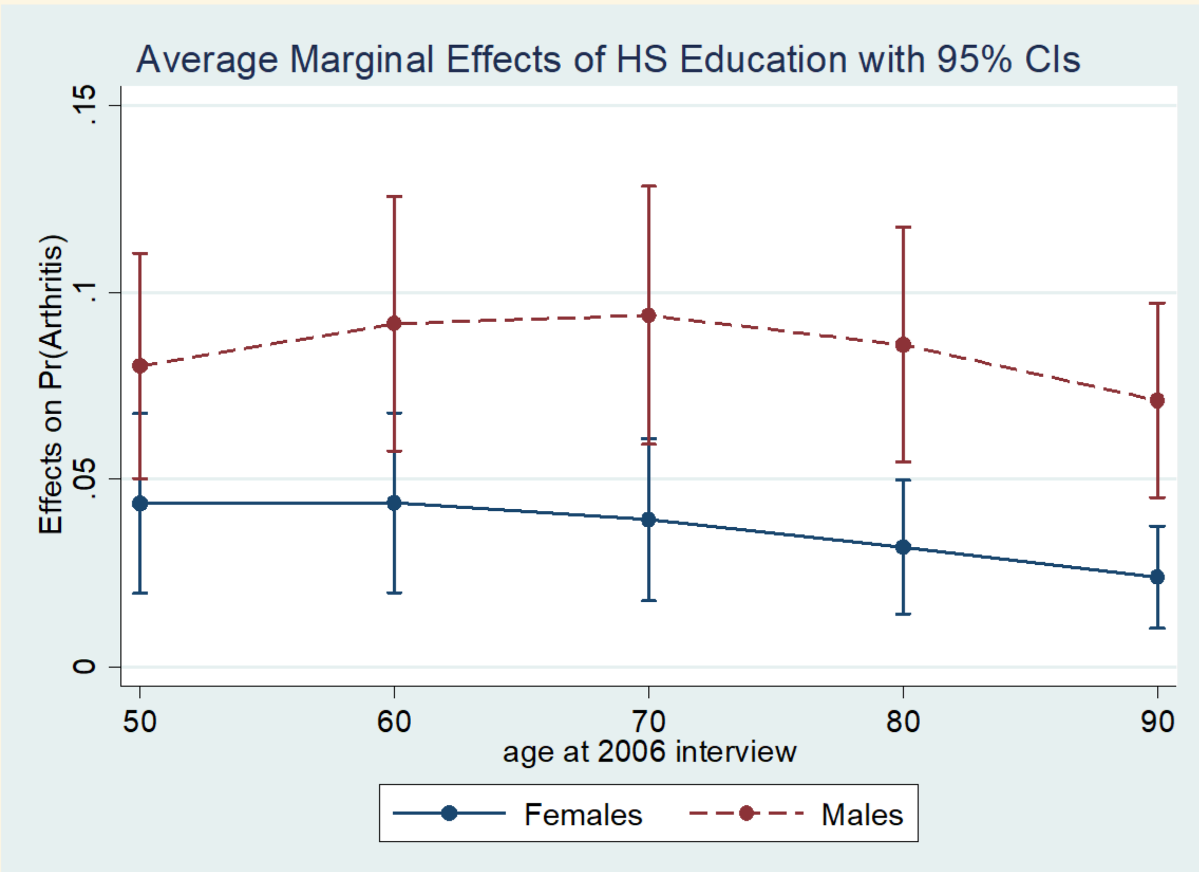
```
svy: logit mde i.ald i.ag4cat
margins ald, by(ag4cat)
marginsplot
```



SOFTWARE PROCEDURES USING STATA

PLOTTING AVERAGE MARGINAL EFFECTS

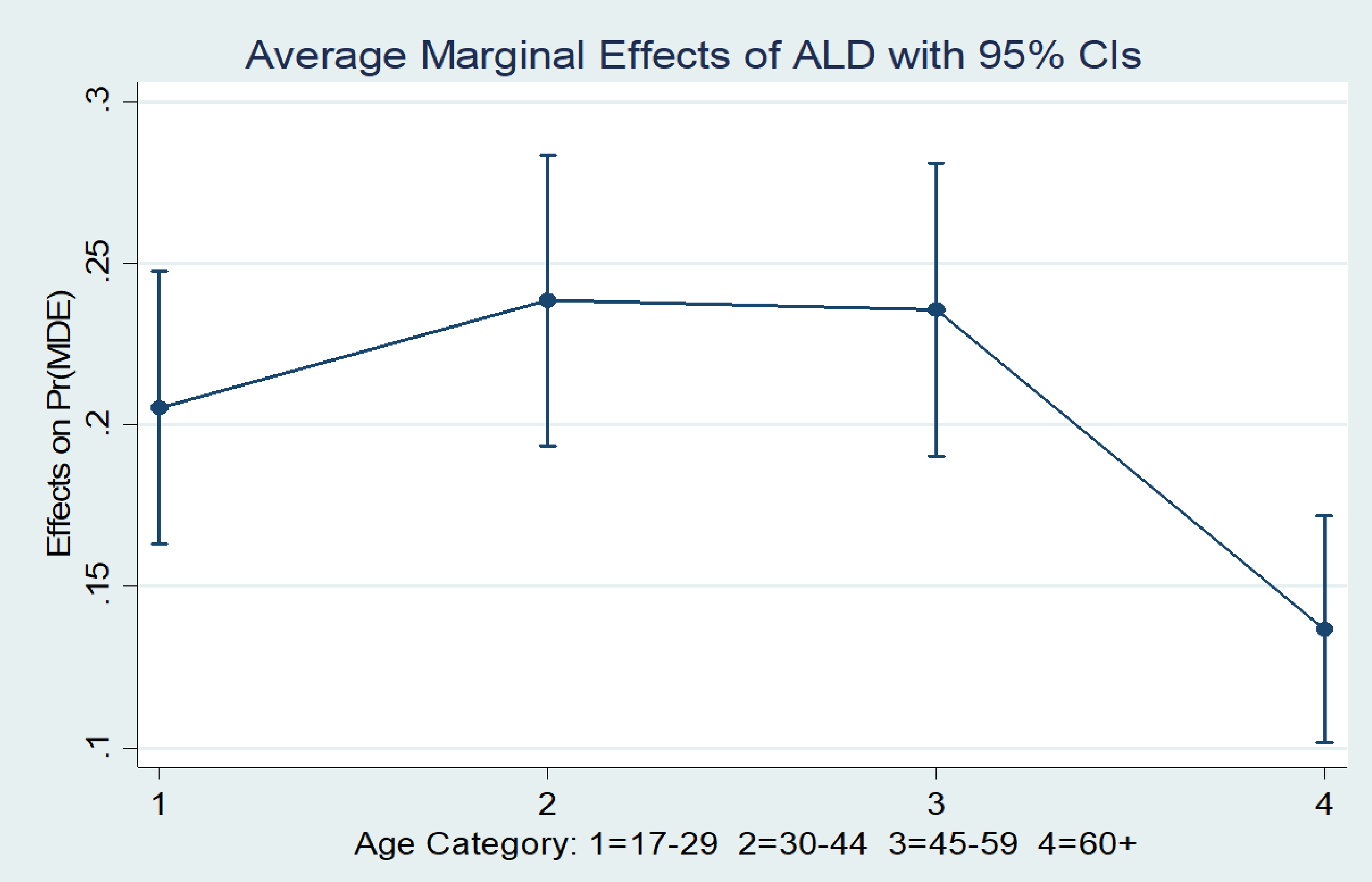
```
margins, dydx(2.edcat3) by(male) at(kage=(50(10)90))
marginsplot
```



SOFTWARE PROCEDURES USING STATA

PLOTTING AVERAGE MARGINAL EFFECTS

```
margins, dydx(ald) by(ag4cat)
marginsplot
```


Average Marginal Effects of ALD with 95% CIs
Effects on Pr(MDE)
.1
.15
.2
.25
.3
1
2
3
4
Age Category: 1=17-29 2=30-44 3=45-59 4=60+