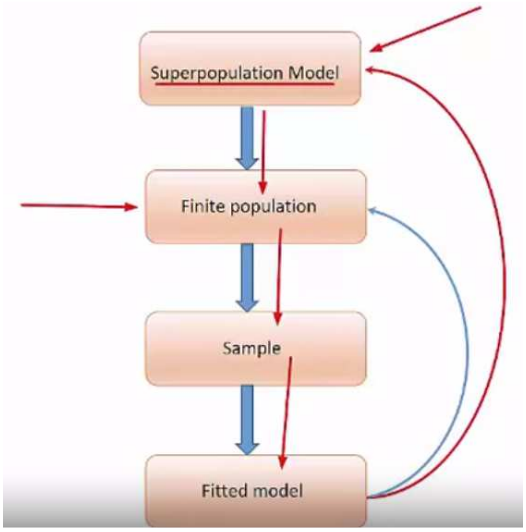
W2 Models

Models



General Considerations
• One way to think of goal of estimation is to estimate the model that would be fitted if entire population were in hand – "census model"
• If same model is appropriate for sample as for full population, then weights do not have to be used
• Stratification and clustering still need to be accounted for because the affect SEs
• But, using weights does insure that we are aiming at the census model

Estimation Method

Pseudo-maximum Likelihood Estimation
• Estimating equation method
• Write down full finite population likelihood
• Derive census estimation equations
  - These will often be finite population totals that involve residuals
• Construct sample estimator of census estimating equations

$$\sum_{i \in s} w_i \mathbf{x}_i (y_i - \mathbf{x}_i^T \hat{\beta}) = 0 \text{ for linear model } y_i = \mathbf{x}_i^T \beta + \epsilon_i$$

• Solve for estimates of model parameters
• Software will do this for quite a few models

Software Capabilities
• R survey
  - linear regression, logistic, probit, complementary log-log, Poisson, Loglinear, Cox proportional hazards model
• Stata
  - linear regression, logistic, probit, complementary log-log, Possion, Loglinear, Cox Proportional hazards model
  - parametric survival, Multinomial logistic, conditional logit, negative binomial, ordered logistic, probit, ordered probit, structural equation modeling, censored and interval regression, instrumental-variables regression, heckman selection model, probit estimation with selection, nonlinear least squares, multilevel models

API dataset in R survey
• Use academic performance index file from R survey
• API is computed for all California schools based on standardized testing of students
• Several datasets: information for all schools with at least 100 students and for various probability samples of data
• One record per school

Variables & Model syntax
• Specify survey design with *svydesign*
• svyglm(formula = …. , design = …..) to fit the model

• Variables
- api00 API in 2000
- ell English language learners (%)
- meals % of students eligible for subsidized meals
- mobility % of students for whom this is the first year at the school

R code

```
require(survey)
data(api)
dstrat <- svydesign(id = ~1, strata = ~stype,
                    weights = ~pw, data = apistrat, fpc = ~fpc)

m1 <- svyglm(api00 ~ ell + meals + mobility, design = dstrat)
summary(m1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 820.8873    10.0777  81.456   <2e-16 ***
ell          -0.4806     0.3920  -1.226    0.222
meals        -3.1415     0.2839 -11.064   <2e-16 ***
mobility      0.2257     0.3932   0.574    0.567
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5171.966)

    Number of Fisher Scoring iterations: 2
```

Testing coefficient estimates

```
regTermTest(m1, ~ell + mobility, method="Wald", df=NULL)

Wald test for ell mobility
 in svyglm(formula = api00 ~ ell + meals + mobility,
        design = dstrat)
F =  1.046306  on  2  and  194  df: p= 0.35321
```

- The joint hypothesis is not rejected that the coefficients are 0 for percentage English language earners and students for whom this is the first year at school

Diagnostics

Diagnostics
• some literature on adapting standard diagnostics for use with survey data
• few options available now in packages ➔ program your own

Compute Standardized residuals
• same model as in previous video
• standardized residuals have mean 0, variance 1 under the model

- Standardized residual is $r_i = (y_i - \hat{y}_i)/\hat{\sigma}$ where

  $\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$ is predicted value

  $\hat{\sigma}^2$ is estimate of variance of error in model (model-variance not design-variance)
- $\hat{\sigma}^2 = \sum_{i \in s} w_i (y_i - \hat{y}_i)^2 / \sum_{i \in s} w_i$

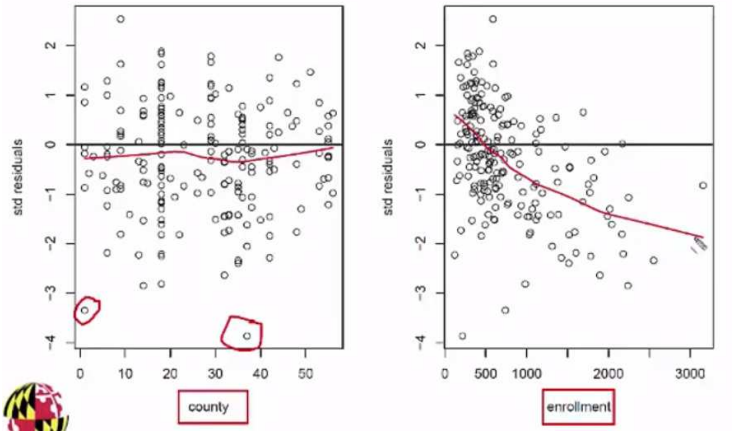
R code to compute standardized residuals

```
require(survey)
data(api)
dstrat <- svydesign(id = ~1, strata = ~stype,
weights = ~pw, data = apistrat, fpc = ~fpc)
m1 <- svyglm(api00 ~ ell + meals + mobility, design = dstrat)
sig2 <- weighted.mean(m1$residuals^2, apistrat$pw)

stdres.m1 <- m1$residuals / sqrt(sig2)
summary(stdres.m1)
       Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    -3.8670 -0.9597 -0.2154 -0.2837  0.5068  2.5430
```

Standardized residual plots

Plotting v.s. variables not in model is a way to look for omitted covariates

```
par(mfrow = c(1,2))
    # plot std residuals vs. county
plot(apistrat$cnum, stdres.m1,
     xlab = "county",
     ylab = "std residuals")
abline(h=0)
lines(lowess(apistrat$cnum, stdres.m1), col="red", lwd=2)

    # plot std residuals vs. enroll
plot(apistrat$enroll, stdres.m1,
     xlab = "enrollment",
     ylab = "std residuals")
abline(h=0)
lines(lowess(apistrat$enroll, stdres.m1), col="red", lwd=2)
```



Linear Models in Stata

Fit same model as in R

• Regress school academic performance indexes (*api00*) on percentage English Language Learners (*ell*), Percentage of students eligible for subsidized meals (*meals*), and percentage of students for whom this is the first year at the school (*mobility*)

```
use apistat.dta, clear
svyset cds [pweight = pw], strata(stype) fpc(fpc)
svy: regress api00 ell meals mobility
```

```
Survey: Linear regression

Number of strata   =         3              Number of obs      =       200
Number of PSUs     =       200              Population size    =     6,194
                                            Design df          =       197
                                            F(   3,    195)    =    135.11
                                            Prob > F           =    0.0000
                                            R-squared          =    0.6595

------------------------------------------------------------------------------
             |             Linearized
      api00 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        ell |  -.4805866   .3919734    -1.23   0.222    -1.253589    .2924159
      meals |  -3.141535   .2839465   -11.06   0.000      -3.7015   -2.58157
  mobility |   .2257132   .3932184     0.57   0.567    -.5497445    1.001171
      _cons |   820.8873   10.07774    81.46   0.000     801.0132    840.7614
------------------------------------------------------------------------------
```

- Same point estimates and SEs as in R
- residuals can be retrieved with `predict r, residuals`
- Standardized residuals need to be computed "by hand" as in R

Test $H_0 : \beta_{ell} = \beta_{mobility} = 0$

```
test (ell=0) (mobility=0)

Adjusted Wald test
 ( 1)  ell = 0
 ( 2)  mobility = 0

       F(  2,   196) =    1.04
            Prob > F =    0.3550
```

- Qualitatively, same result as in R: do not reject
- Note that denominator *df* = 196 not 194. Slightly different adjustment based on number of parameters tested
- Difference is not important here

Logistic Models in R

Logistic Example using API dataset

• Logistic model to predict whether a school met target for school-wide growth in API score

• Test whether subset of coefficients in 0

• Odds ratio for a categorical predictor

API Dataset

Variable	Description
`sch.wide`	Met school-wide growth target? (N or Y)
`ell`	Percentage of English language learners
`meals`	Percentage of students eligible for subsidized meals
`mobility`	Percentage of students for whom this is the first year at the school
`enroll`	Number of students enrolled
`hsg`	Percentage of parents who are high-school graduates
`col.grad`	Percentage of parents with college degree
`yr.rnd`	Year-round school (N or Y)

R survey code

```
require(survey)
data(api)
dstrat <- svydesign(id = ~1, strata = ~stype,
                    weights = ~pw, data = apistrat, fpc = ~fpc)

m2 <- svyglm(sch.wide ~ ell + meals + mobility + enroll + hsg + col.grad
        design=dstrat, family = quasibinomial(link = logit))
summary(m2)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.3788723  0.9322200   1.479   0.1408
ell                -0.0047223  0.0157843  -0.299   0.7651
meals               0.0006536  0.0133165   0.049   0.9609
mobility            0.0379475  0.0215958   1.757   0.0805 .
enroll             -0.0020832  0.0003898  -5.344 2.59e-07 ***
hsg                 0.0086906  0.0131065   0.663   0.5081
col.grad            0.0395304  0.0230179   1.717   0.0875 .
factor(yr.rnd)Yes   1.4164830  0.8995009   1.575   0.1170
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test a subset of coefficients

```
regTermTest(m2, ~ell + meals + hsg, method="Wald", df=NULL)
Wald test for ell meals hsg
 in svyglm(formula = sch.wide ~ ell + meals + mobility +
    enroll + hsg + col.grad + factor(yr.rnd), design = dstrat,
    family = quasibinomial(link = logit))
F =  0.2395591  on  3  and  190  df: p= 0.86868
```

- Cannot reject joint hypothesis that `ell, meals, hsg` are all 0

Odds Ratios

• The odds of having a characteristic is p/(1-p)

• The ratio of the odds (odds ratio) of having a characteristic for category 1 of a predictor to category 0 is

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

- The logistic model is $log[\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}] = \mathbf{x}_i^T\beta$ where $p(\mathbf{x}_i)$ depends on the covariates for unit i

(LOG of the odds)

Transforming parameter estimates to ORs

- $log(OR)$ for a unit in category 1 vs. a unit in category 0 of a covariate, setting all other covariates the same for the two units is

$$log[p_1/(1-p_1)] - log[p_0/(1-p_0)] = \beta_1 - \beta_0$$

- In a logistic regression, if level 0 is the reference category, then $\hat{\beta}_0$ is set to 0
- Transform to OR scale as $\widehat{OR} = exp(\hat{\beta}_1)$

R survey code

```
require(survey)
data(api)
dstrat <- svydesign(id = ~1, strata = ~stype,
                    weights = ~pw, data = apistrat, fpc = ~fpc)
m3 <- svyglm(sch.wide ~ mobility + enroll + col.grad + factor(yr.rnd),
        design=dstrat, family = quasibinomial(link = logit))
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.4674032  0.5965818   2.460   0.0148 *
mobility            0.0395838  0.0219135   1.806   0.0724 .
enroll             -0.0020771  0.0003934  -5.280 3.46e-07 ***
col.grad            0.0399130  0.0176777   2.258   0.0251 *
factor(yr.rnd)Yes   1.3241534  0.9086414   1.457   0.1467

exp(1.3241534)
[1] 3.759002
```

- Year-round schools 3.76 times more likely to meet school-wide growth target

Confidence Intervals

```
CI <- confint(m3)
                              2.5 %       97.5 %
(Intercept)            0.29812436l  2.636682062
mobility              -0.003365872  0.082533500
enroll                -0.002848080 -0.001306086
col.grad               0.005265443  0.074560559
factor(yr.rnd)Yes -0.456751056  3.105057838

exp(CI[5,])
    2.5 %    97.5 %
 0.633338 22.310509
```

- 95% CI for odds ratio of Year-round school vs. Not Year-round covers 1
- Point estimate of OR is suggestive that year-round schools are more likely to hit target but estimate is not precise enough to be sure

Logistic Regression in Stata

Fit same model as in R

• Regress school-wide growth target met (sch.wide) on enrollment (enroll), percentage of parents with college degree (col.grad), and indicator for whether school is year-round or not (yr.rnd)

```
use apistrat.dta, clear
svyset cds [pweight = pw], strata(stype) fpc(fpc)
    * recode sch_wide
gen sw01 = sch_wide
recode sw01 (1 = 0) (2 = 1)
    * logistic with coefficients
svy: logit sw01 mobility enroll col_grad i.yr_rnd
Number of strata   =         3          Number of obs      =         200
Number of PSUs     =       200          Population size    =       6,194
                                        Design df          =         197
                                        F(   4,     194)   =        8.83
                                        Prob > F           =      0.0000
------------------------------------------------------------------------------
             |             Linearized
        sw01 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mobility |   .0395838   .0219137     1.81   0.072    -.0036318    .0827994
      enroll |  -.0020771   .0003934    -5.28   0.000    -.0028528   -.0013013
    col_grad |    .039913   .0176776     2.26   0.025     .0050513    .0747747
      yr_rnd |
         Yes |   1.324153   .9086466     1.46   0.147    -.4677695    3.116076
       _cons |   1.467403   .5965847     2.46   0.015     .2908911    2.643915
------------------------------------------------------------------------------
```

Logit on odds ratios

```
    * logistic with odds ratios
svy: logistic sw01 mobility enroll col_grad i.yr_rnd
------------------------------------------------------------------------------
             |             Linearized
        sw01 | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mobility |   1.040378   .0227986     1.81   0.072     .9963748    1.086324
      enroll |   .9979251   .0003926    -5.28   0.000     .9971512    .9986995
    col_grad |    1.04072   .0183975     2.26   0.025     1.005064    1.077641
             |
      yr_rnd |
         Yes |   3.759002   3.415604     1.46   0.147     .6263979     22.5577
       _cons |   4.337956   2.587958     2.46   0.015     1.337619    14.06818
------------------------------------------------------------------------------
```

- Same regression coefficient estimates and OR's as R `survey` for `yr_rnd Yes`
- Slightly different SE's and CI's for OR's