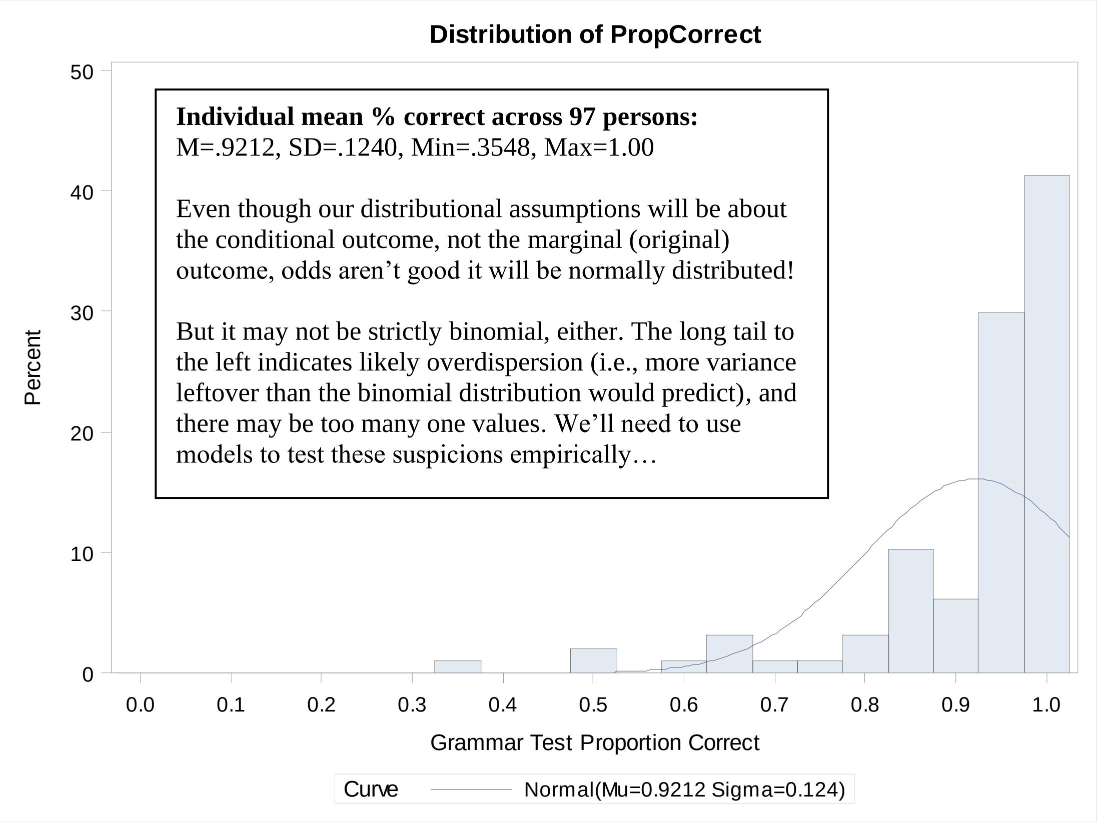
Example 4a: Generalized Linear Models for Binomial Outcomes (% Correct)
(complete syntax and output available for SAS, STATA, and R electronically)

The real data for this example come from the publication below, which examined annual growth in a test of grammatical understanding from Kindergarten through 4th grade in children with non-specific language impairment (NLI) or specific language impairment (SLI):

Rice, M. L., Tomblin, J. B., **Hoffman, L.**, Richman, W. A., & Marquis, J. (2004). Grammatical tense deficits in children with SLI and nonspecific language impairment: Relationships with nonverbal IQ over time. *Journal of Speech-Language-Hearing Research, 47*(4), 816-834.

The current example is a cross-sectional analysis of how grammatical understanding at third grade (measured by proportion correct) is predicted by group (NLI=0, SLI=1) and mother's years of education (centered so that 0=12 years). Given that proportion correct is bounded by 0 and 1, we will use a logit link and the binomial family of conditional response distributions. Because the binomial is a discrete distribution, we will need to parameterize the model to predict the number of correct responses out of the number of trials directly instead of proportion correct. This example will also demonstrate two ways of addressing binomial overdispersion: *additive* (through individual random intercepts) and *multiplicative* (through the beta-binomial distribution), as well as zero-inflated (actually one-inflated here; stay tuned) versions of the binomial and beta-binomial model variants (in which the probability of being an extra zero is predicted in a separate submodel using a logit link).

In SAS, I am still using GLIMMIX for the binomial models, as well as FMM (finite mixture model) for the beta-binomial and zero-inflated model variants. Further, because the relevant STATA options (using GLM to get conditional distribution fit, also using MEGLM, BETABIN, ZIB, and ZIBBIN here) do not have denominator degrees of freedom, they were set to "none" in SAS GLIMMIX so that the SAS Wald test results (still labeled as t or F) will match those of STATA (using z or χ^2). In R, I am using VGLM from the VGAM package and GLMER from the LME4 package (each also using z or χ^2). There are some inconsistencies in the results from R (as usual) where noted.



Individual mean % correct across 97 persons:
M=.9212, SD=.1240, Min=.3548, Max=1.00

Even though our distributional assumptions will be about the conditional outcome, not the marginal (original) outcome, odds aren't good it will be normally distributed!

But it may not be strictly binomial, either. The long tail to the left indicates likely overdispersion (i.e., more variance leftover than the binomial distribution would predict), and there may be too many one values. We'll need to use models to test these suspicions empirically…

SAS Data Manipulation and Description:

```
* Location for original files for these models - change this path;
* \\Client\ precedes path in Virtual Desktop outside H drive;
%LET filesave=C:\Dropbox\22_PSQF6270\PSQF6270_Example4a;
LIBNAME filesave "&filesave.";

* Import Example 4a SAS data into work library;
DATA work.Example4a; SET filesave.PSQF6270_Example4a;
* Label existing variables for analysis;
  LABEL NLIvSLI=     "Group: 0=NLI, 1=SLI"
        momed12=     "Mother Education (0=12 years")
        PropCorrect= "Grammar Test Proportion Correct";
* Create number correct for denominator of binomial outcome;
  Ntrials=100;
  Ncorrect=ROUND(PropCorrect*Ntrials,1);
* Compute number incorrect for zero-inflated binomial model;
  Nincorrect=Ntrials-Ncorrect;
  PropIncorrect=1-PropCorrect;
RUN;

TITLE "SAS Distribution of Proportion Correct";
PROC MEANS NDEC=3 DATA=work.Example4a;
     VAR PropCorrect;
RUN;
PROC UNIVARIATE NOPRINT DATA=work.Example4a;
     VAR PropCorrect;
     HISTOGRAM PropCorrect / MIDPOINTS= 0 TO 1 BY .05 NORMAL(MU=EST SIGMA=EST);
RUN; QUIT; TITLE;
```

STATA Data Manipulation and Description:

```
// Defining global variable for file location to be replaced in code below
// \\Client\ precedes path in Virtual Desktop outside H drive;
global filesave "C:\Dropbox\22_PSQF6270\PSQF6270_Example4a"

// Import Example 4a Stata data
use "$filesave\PSQF6270_Example4a.dta", clear

// Label existing vriables for analysis
label variable nlivsli     "Group: 0=NLI, 1=SLI"
label variable momed12     "Mother Education (0=12 years)"
label variable propcorrect "Grammar Test Proportion Correct"
// Create number correct for denominator of binomial outcome
gen ntrials=100
gen ncorrect=round(propcorrect*ntrials,1)
// Compute number incorrect for zero-inflated binomial model
gen nincorrect=ntrials-ncorrect
gen propincorrect=1-propcorrect

// Find betabin and zbin, install before continuing
// search betabin
// search zbin

display "STATA Distribution of Proportion Correct"
summarize propcorrect
hist propcorrect, percent start(0) width(.05)
graph export "$filesave\STATA Proportion Correct Histogram.png", replace
```

R Data Manipulation and Description:

```
# Define variables for working directory and data name
filesave = "C:\\Dropbox/22_PSQF6270/PSQF6270_Example4a/"
filename = "PSQF6270_Example4a.sas7bdat"
setwd(dir=filesave)
```

```
# Import Example 4a SAS data
Example4a = read_sas(data_file=paste0(filesave,filename))
# Convert to data frame without labels to use for analysis
Example4a = as.data.frame(Example4a)

# Label existing variables for analysis
#NLIvSLI=     "Group: 0=NLI, 1=SLI"
#momed12=     "Mother Education (0=12 years")
#PropCorrect= "Grammar Test Proportion Correct"

# Create number correct for denominator of binomial outcome
Example4a$Ntrials=100
Example4a$Ncorrect=Example4a$PropCorrect*Example4a$Ntrials
Example4a$Ncorrect=round(Example4a$Ncorrect,0)
# Compute number incorrect for zero-inflated binomial model
Example4a$Nincorrect=Example4a$Ntrials-Example4a$Ncorrect
Example4a$PropIncorrect=1-Example4a$PropCorrect

print("R Distribution of Proportion Correct")
describe(x=Example4a$PropCorrect)

# to save a plot: open a file, create the plot, then close the file
png(file = "R Proportion Correct Histogram.png")  # open file
hist(x=Example4a$PropCorrect, freq=FALSE,
     ylab="Density",xlab="Grammar Test Proportion Correct") # axis labels
dev.off()  # close file
```

---

1) Empty Binomial Model for % correct
DV = Events/Trials in SAS and STATA; Events/Non-Events in R

$\#Correct_i \sim Binomial(p_i, Ntrials_i)$ → p_i is probability of any one trial being correct

$Logit(p_i \text{ for correct trial}) = \beta_0$

Conditional mean for $\#Correct_i \ = Ntrials_i * p_i$

Conditional variance for $\#Correct_i = (Ntrials_i * p_i)(1 - p_i)$

```
TITLE "SAS Empty Binomial Model using GLIMMIX -- Ntrials is denominator";
PROC GLIMMIX DATA=work.Example4a NOCLPRINT NAMELEN=100 GRADIENT METHOD=MSPL;
     MODEL Ncorrect/Ntrials = / SOLUTION DDFM=NONE LINK=LOGIT DIST=BINOMIAL;
     ESTIMATE "Intercept" intercept 1 / ILINK; * ILINK gives intercept in probability;
RUN; TITLE;
```

```
display "STATA Empty Binomial Model using glm -- ntrials is denominator"
glm ncorrect, link(logit) family(binomial ntrials)
display "-2LL=" e(ll)*-2  // Print -2LL for model
estat ic, n(97)  // AIC and BIC matches SAS
margins          // Get intercept in percent (ILINK*Ntrials)
```

```
print("R Empty Binomial Model using vglm and two outcome columns")
ModelEmpty = vglm(data=Example4a, binomialff(link="logitlink", multiple.responses=FALSE),
                  formula=cbind(Ncorrect,Nincorrect)~1) # Can also use multiv format of 0/1
summary(ModelEmpty)
print("Get intercept in probability")
ModelEmptyProb=1/(1+exp(-1*coefficients(ModelEmpty))); ModelEmptyProb
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelEmpty); AIC(ModelEmpty); BIC(ModelEmpty)
print("Pearson Chi-Square / DF Index of Fit matching SAS and STATA")
sum(residuals(ModelEmpty, type="pearson")^2)/(97)    # SAS N
sum(residuals(ModelEmpty, type="pearson")^2)/(97-1)  # STATA N-k
```

Partial SAS Output:

```
           Fit Statistics
-2 Log Likelihood              1841.19
AIC  (smaller is better)       1843.19
BIC  (smaller is better)       1845.76
Pearson Chi-Square             2041.44
Pearson Chi-Square / DF          21.05
```

To inverse link from logits to predicted % correct:

$$\text{Prob}(y=1) = \frac{\exp(2.4593)}{1+\exp(2.4593)} = .9212$$

The sample average probability of getting each item correct is .9212.

But Chi-Square/DF > 1, indicating that this model has over-dispersion (too much variance, likely in part because we haven't incorporated predictors yet).

```
                             Parameter Estimates
                         Standard
Effect       Estimate       Error       DF    t Value    Pr > |t|    Gradient
Intercept      2.4593     0.03769    Infty      65.24      <.0001    -1.03E-6

                                        Estimates
                         Standard                                                Standard Error
Label        Estimate       Error       DF    t Value    Pr > |t|        Mean          Mean
Intercept      2.4593     0.03769    Infty      65.24      <.0001      0.9212      0.002735
```

So even though we are actually modeling number of correct trials as the DV, the model is phrased to **predict proportion correct directly** (as the conditional mean *p*, the probability that any trial = 1).

Partial STATA Output:

```
Generalized linear models                          No. of obs      =         97
Optimization     : ML                              Residual df     =         96
                                                   Scale parameter =          1
Deviance         =   1620.05009                    (1/df) Deviance =   16.87552
Pearson          =  2041.435988                    (1/df) Pearson  =   21.26496 → / N-k instead
Variance function: V(u) = u*(1-u/ntrials)          [Binomial]
Link function    : g(u) = ln(u/(ntrials-u))        [Logit]
                                                   AIC             =   19.00196 → not usual version!
Log likelihood   = -920.5951086                    BIC             =   1180.878 → not usual version!
------------------------------------------------------------------------------
             |                 OIM
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   2.459276   .0376936    65.24   0.000     2.385397    2.533154
------------------------------------------------------------------------------
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   92.12371   .2735021   336.83   0.000     91.58766    92.65977 → intercept in percent
------------------------------------------------------------------------------
```

Partial R Output:

```
Coefficients:
            Estimate Std. Error z value              Pr(>|z|)
 (Intercept) 2.459276   0.037693  65.246 < 0.0000000000000022

Name of linear predictor: logitlink(prob)
Residual deviance: 1620.0501 on 96 degrees of freedom
Log-likelihood: -920.59511 on 96 degrees of freedom

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept)'

[1] "Get intercept in probability"
> ModelEmptyProb = 1/(1 + exp(-1 * coefficients(ModelEmpty)))
 0.92123711

[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelEmpty)
[1] 1841.1902
> AIC(ModelEmpty)
```

This warning indicates that the intercept had some difficulty being estimated. For more info, see https://search.r-project.org/CRAN/refmans/VGAM/html/hdeff.html

```
[1] 1843.1902
> BIC(ModelEmpty)
[1] 1845.7649

[1] "Pearson Chi-Square / DF Index of Fit matching SAS and STATA"
> sum(residuals(ModelEmpty, type = "pearson")^2)/(97)
[1] 21.043643
> sum(residuals(ModelEmpty, type = "pearson")^2)/(97 - 1)
[1] 21.262847
```

---

2) Two-Predictor Binomial Model

$\#Correct_i \sim Binomial(p_i, Ntrials_i)$ → p_i is probability of any one trial being correct

$Logit(p_i$ for correct trial$) = \beta_0 + \beta_1(NLIvSLI_i) + \beta_2(MotherEd_i - 12)$

Conditional mean: $\#Correct_i \ = Ntrials_i * p_i$

Conditional variance of $\#Correct_i = (Ntrials_i * p_i)(1 - p_i)$

```
TITLE "SAS Two-Predictor Binomial Model using GLIMMIX";
PROC GLIMMIX DATA=work.Example4a NOCLPRINT NAMELEN=100 GRADIENT METHOD=MSPL;
     MODEL Ncorrect/Ntrials = NLIvSLI momed12
            / SOLUTION DDFM=NONE LINK=LOGIT DIST=BINOMIAL ODDSRATIO(LABEL);
     CONTRAST "Multiv Wald test of Model" NLIvSLI 1, momed12 1  / CHISQ;
RUN; TITLE;

display "STATA Two-Predictor Binomial Model using glm"
glm ncorrect c.nlivsli c.momed12, link(logit) family(binomial ntrials)
display "-2LL=" e(ll)*-2  // Print -2LL for model
estat ic, n(97)  // AIC and BIC matches SAS
test (c.nlivsli=0)(c.momed12=0)  // Multiv Wald test of model

display "STATA Two-Predictor Binomial Model -- get odds ratios using eform"
glm ncorrect c.nlivsli c.momed12, link(logit) family(binomial ntrials) eform

print("R Two-Predictor Binomial Model using vglm")
print("Parameter SEs do not match SAS and STATA exactly but are close")
ModelBin = vglm(data=Example4a, binomialff(link="logitlink", multiple.responses=FALSE),
                formula=cbind(Ncorrect,Nincorrect)~1+NLIvSLI+momed12)
summary(ModelBin)
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelBin); AIC(ModelBin); BIC(ModelBin)
print("Pearson Chi-Square / DF Index of Fit matching SAS and STATA")
sum(residuals(ModelBin, type="pearson")^2)/(97)    # SAS N
sum(residuals(ModelBin, type="pearson")^2)/(97-3)  # STATA N-k

print("Multiv Wald test of model -- very close to SAS and STATA")
BinR2 = glht(model=ModelBin, linfct=c("NLIvSLI=0","momed12=0"))
print(summary(BinR2, test=Chisqtest()), digits="8") # Joint chi-square test
```

Partial SAS Output:

```
           Fit Statistics
-2 Log Likelihood            1531.73
AIC  (smaller is better)     1537.73
BIC  (smaller is better)     1545.46
Pearson Chi-Square           1448.89
Pearson Chi-Square / DF        14.94 → better, but nowhere good enough!

                            Parameter Estimates
                        Standard
Effect       Estimate      Error       DF    t Value    Pr > |t|    Gradient
Intercept      3.0719    0.07462    Infty      41.17      <.0001    -5.77E-6  Beta0 → proportion=.956
NLIvSLI       -1.2216    0.08587    Infty     -14.23      <.0001    -3.06E-9  Beta1
momed12        0.1193    0.02143    Infty       5.57      <.0001    -6.69E-6  Beta2
```

```
                              Odds Ratio Estimates
                                                           95% Confidence
Comparison                            Estimate       DF          Limits
unit change of NLIvSLI from mean        0.295    Infty       0.249       0.349
unit change of momed12 from mean        1.127    Infty       1.080       1.175

                                        Contrasts
Label                       Num DF   Den DF    Chi-Square    F Value      Pr > ChiSq    Pr > F
Multiv Wald test of Model        2    Infty        273.58     136.79          <.0001    <.0001
```

Partial STATA Output:

```
Generalized linear models                          No. of obs      =         97
Optimization     : ML                              Residual df     =         94
                                                   Scale parameter =          1
Deviance         =  1310.593044                    (1/df) Deviance =   13.94248
Pearson          =  1448.891028                    (1/df) Pearson  =   15.41373 → / N-k instead
Variance function: V(u) = u*(1-u/ntrials)          [Binomial]
Link function    : g(u) = ln(u/(ntrials-u))        [Logit]
                                                   AIC             =   15.85292
Log likelihood   = -765.8665858                    BIC             =   880.5702
------------------------------------------------------------------------------
             |                 OIM
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nlivsli |  -1.221578   .0858707   -14.23   0.000    -1.389881   -1.053275  Beta1
     momed12 |   .1193325   .0214268     5.57   0.000     .0773368    .1613283  Beta2
       _cons |   3.071929   .0746183    41.17   0.000      2.92568    3.218178  Beta0 → proportion=.956
------------------------------------------------------------------------------
. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =  273.58
         Prob > chi2 =    0.0000
```

Partial R Output:

```
Coefficients:
             Estimate Std. Error  z value              Pr(>|z|)
(Intercept)  3.071929   0.074610  41.1729 < 0.00000000000000022  Beta0 → proportion=.956
NLIvSLI     -1.221578   0.085864 -14.2269 < 0.00000000000000022  Beta1
momed12      0.119333   0.021426   5.5695         0.00000002555  Beta2

Name of linear predictor: logitlink(prob)
Residual deviance: 1310.593 on 94 degrees of freedom
Log-likelihood: -765.86659 on 94 degrees of freedom

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept)'

[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelBin)
[1] 1531.7332
> AIC(ModelBin)
[1] 1537.7332
> BIC(ModelBin)
[1] 1545.4573

[1] "Pearson Chi-Square / DF Index of Fit matching SAS and STATA"
> sum(residuals(ModelBin, type = "pearson")^2)/(97)
[1] 14.934176
> sum(residuals(ModelBin, type = "pearson")^2)/(97 - 3)
[1] 15.410798

[1] "Multiv Wald test of model -- very close to SAS and STATA"

         General Linear Hypotheses
Linear Hypotheses:
                Estimate
NLIvSLI == 0 -1.22157804
momed12 == 0  0.11933255

Global Test:
      Chisq DF    Pr(>Chisq)
1 273.62108  2 3.8365119e-60
```

Before interpreting these results, let's see if we can get better distribution fit. Here are some alternative models that incorporate overdispersion, zero-inflation (actually one-inflation here), or both at the same time…

3) Two-Predictor Binomial Model with Additive Overdispersion: Also known as adding an "observation-level random effect" (OLRE)

$\#Correct_i \sim Binomial(p_i, Ntrials_i)$ → p_i is probability of any one trial being correct

$Logit(p_i \text{ for correct}) = \beta_0 + \beta_1(NLIvSLI_i) + \beta_2(MotherEd_i - 12) + U_{0i}$

Conditional mean of $\#Correct_i \ = Ntrials_i * p_i$

Conditional variance of $\#Correct_i = (Ntrials_i * p_i)(1 - p_i)$

The random intercept variance is on the model-scale (in logits), and it effectively soaks up all discrepancy to each person's predicted logit.

```
TITLE1 "SAS Two-Predictor Binomial Model using GLIMMIX";
TITLE2 "Additive Overdispersion via Random Intercept Variance";
TITLE3 "Also known as observation-level random effect for overdispersion";
PROC GLIMMIX DATA=work.Example4a NOCLPRINT NAMELEN=100 GRADIENT METHOD=LAPLACE;
     CLASS ID;  * Person ID is added to CLASS because of RANDOM statement below;
     MODEL Ncorrect/Ntrials = NLIvSLI momed12
             / SOLUTION DDFM=NONE LINK=LOGIT DIST=BINOMIAL ODDSRATIO(LABEL);
     CONTRAST "Multiv Wald test of Model" NLIvSLI 1, momed12 1 / CHISQ;
     RANDOM INTERCEPT / SUBJECT=ID;   * Add per-person "residual" as random intercept;
     COVTEST "Need Random Intercept Variance?" 0;  * LRT for additive overdispersion;
RUN; TITLE1; TITLE2; TITLE3;
```

```
display "STATA Two-Predictor Binomial Model with Additive Overdispersion"
display "Using meglm instead; || id. adds random intercept variance"
display "Also known as observation-level random effect for overdispersion"
meglm ncorrect c.nlivsli c.momed12, || id: , ///
      link(logit) family(binomial ntrials) intmethod(laplace)
display "-2LL=" e(ll)*-2  // Print -2LL for model
estat ic, n(97)  // AIC and BIC matches SAS
test (c.nlivsli=0)(c.momed12=0) // Multiv Wald test of model (given in meglm)
// LRT for added random intercept variance is done for you automatically

display "STATA Two-Predictor Binomial Model with Additive Overdispersion"
display "Using meglm instead and getting odds ratios using eform"
meglm ncorrect c.nlivsli c.momed12, || id: , ///
      link(logit) family(binomial ntrials) intmethod(laplace) eform
```

```
print("R Two-Predictor Binomial Model using glmer")
print("Additive Overdispersion via Random Intercept Variance (1|ID)")
print("Also known as observation-level random effect for overdispersion")
print("Parameter SEs do not match SAS and STATA exactly but are close")
ModelBinAdd = glmer(data=Example4a, family=binomial(link="logit"),
                    formula=cbind(Ncorrect,Nincorrect)~1+NLIvSLI+momed12+(1|ID))
summary(ModelBinAdd)
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelBinAdd); AIC(ModelBinAdd); BIC(ModelBinAdd)
print("Pearson Chi-Square / DF Index of Fit matching SAS and STATA")
sum(residuals(ModelBinAdd, type="pearson")^2)/(97)  # SAS N

print("Multiv Wald test of model -- very close to SAS and STATA")
BinAddR2 = glht(model=ModelBinAdd, linfct=c("NLIvSLI=0","momed12=0"))
print(summary(BinAddR2, test=Chisqtest()), digits="8") # Joint chi-square test

print("Likelihood Ratio Test for Addition of Random Intercept Variance")
DevTestA=-2*(logLik(ModelBin)-logLik(ModelBinAdd))
RegPvalueA=pchisq((DevTestA), df=1, lower.tail=FALSE)
MixPvalueA=RegPvalueA/2
print("Test Statistic, Regular and Mixture P-values for DF=1")
DevTestA; RegPvalueA; MixPvalueA
```

Partial SAS Output:

```
           Fit Statistics
-2 Log Likelihood              549.76
AIC  (smaller is better)       557.76
BIC  (smaller is better)       568.10

Fit Statistics for Conditional Distribution
-2 log L(Ncorrect | r. effects)     248.93
Pearson Chi-Square                   14.47
Pearson Chi-Square / DF               0.15 → Much lower because extra variance is included in the model

             Covariance Parameter Estimates
                                   Standard
Cov Parm     Subject    Estimate      Error    Gradient
Intercept    ID           4.3810     1.0287    -0.00019 → Extra variance on logit model-scale

                         Solutions for Fixed Effects
                         Standard
Effect       Estimate       Error       DF    t Value    Pr > |t|    Gradient
Intercept      4.7427      0.4351    Infty      10.90      <.0001    0.000149 → proportion=.991
NLIvSLI       -1.7937      0.5089    Infty      -3.52      0.0004    -0.00016
momed12       0.03278      0.1341    Infty       0.24      0.8069    -0.00099 → no longer significant

                                             Contrasts
Label                        Num DF   Den DF    Chi-Square    F Value   Pr > ChiSq    Pr > F
Multiv Wald test of Model         2    Infty         14.04       7.02       0.0009    0.0009 → big difference!

                        Tests of Covariance Parameters
                           Based on the Likelihood
Label                         DF    -2 Log Like      ChiSq    Pr > ChiSq    Note
Need Extra Variance?           1        1531.73     981.97        <.0001    MI → LRT with mixture of DF=0,1
```

Along with the much larger standard errors (as expected from allowing extra variance), the estimates have also changed because the total model has more variance in it now (as opposed to the total 3.29 residual variance given the logit link).

Partial STATA Output:

```
Mixed-effects GLM                               Number of obs     =         97
Family:                 binomial
Link:                      logit
Binomial variable:      ntrials
Group variable:              id                 Number of groups  =         97

                                                Obs per group:
                                                              min =          1
Integration method:     laplace                               avg =        1.0
                                                              max =          1
                                                Wald chi2(2)      =      14.04
Log likelihood = -274.88176                     Prob > chi2       =     0.0009
------------------------------------------------------------------------------
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nlivsli |  -1.793682   .5089051    -3.52   0.000    -2.791118   -.7962467
     momed12 |   .0327918   .1341283     0.24   0.807    -.2300949    .2956784
       _cons |   4.742648   .4350784    10.90   0.000      3.88991    5.595386 → proportion=.991
-------------+----------------------------------------------------------------
id           |
   var(_cons)|   4.381075   1.028769                      2.765034    6.941619 → extra variance on logit scale
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 981.97      Prob >= chibar2 = 0.0000 → LRT of additive overdispersion

. test (c.nlivssli=0)(c.momed12=0) // Multiv Wald test of model
           chi2(  2) =   14.04
         Prob > chi2 =    0.0009
```

Partial R Output:

```
     AIC      BIC   logLik deviance df.resid
   557.8    568.1   -274.9    549.8       93 → btw, deviance is -2LL

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 4.3795   2.0927
Number of obs: 97, groups:  ID, 97

Fixed effects:
             Estimate Std. Error z value              Pr(>|z|)
(Intercept)  4.741983   0.433979 10.9268 < 0.00000000000000022 → proportion=.991
NLIvSLI     -1.793161   0.508270 -3.5280             0.0004188
momed12      0.032825   0.134020  0.2449             0.8065140

[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelBinAdd)
'log Lik.' 549.76432 (df=4)
> AIC(ModelBinAdd)
[1] 557.76432
> BIC(ModelBinAdd)
[1] 568.06316

[1] "Pearson Chi-Square / DF Index of Fit matching SAS and STATA"
> sum(residuals(ModelBinAdd, type = "pearson")^2)/(97)
[1] 0.14926008

[1] "Multiv Wald test of model -- very close to SAS and STATA"
        General Linear Hypotheses
Linear Hypotheses:
                 Estimate
NLIvSLI == 0 -1.793161314
momed12 == 0  0.032824995

Global Test:
      Chisq DF    Pr(>Chisq)
1 14.061888  2 0.00088409677

[1] "Likelihood Ratio Test for Addition of Random Intercept Variance"
> DevTestA = -2 * (logLik(ModelBin) - logLik(ModelBinAdd))
> RegPvalueA = pchisq((DevTestA), df = 1, lower.tail = FALSE)
> MixPvalueA = RegPvalueA/2

[1] "Test Statistic, Regular and Mixture P-values for DF=1"
> DevTestA
'log Lik.' 981.96886 (df=4)
> RegPvalueA
'log Lik.' 1.4914969e-215 (df=4)
> MixPvalueA
'log Lik.' 7.4574845e-216 (df=4)
```

---

4) Two-Predictor Model with Multiplicative Overdispersion via the Beta-Binomial Distribution

$\#Correct_i \sim BetaBinomial(p_i, Ntrials_i, \phi)$ → p_i is still probability of any one trial being correct

$p_i \sim Beta(a_i, b_i)$ → $a_i = p_i/\phi,\ b_i = (1-p_i)/\phi$

$Logit(p_i \text{ for correct trial}) = \beta_0 + \beta_1(NLIvSLI_i) + \beta_2(MotherEd_i - 12)$

Conditional mean: $\#Correct_i = Ntrials_i * p_i$

Conditional variance of $\#Correct_i = (Ntrials_i * p_i)(1-p_i)[1 + (Ntrials_i - 1)/(\phi + 1)]$

Disclaimer: I struggled to translate this model across the different parameterizations I found, and this formula for the conditional variance produced results that were close to those of SAS but not exactly the same…

```
TITLE1 "SAS Two-Predictor Beta-Binomial Model with Multiplicative Overdispersion";
TITLE2 "Using PROC FINITE MIXTURE MODEL that has beta-binomial distribution";
PROC FMM DATA=work.Example4a NAMELEN=100;
     MODEL Ncorrect/Ntrials = NLIvSLI momed12 / LINK=LOGIT DIST=BETABINOMIAL;
     OUTPUT OUT=work.BBpred PRED=yhat VAR=yhatvar;
RUN; TITLE1; TITLE2;
```

```
* Compute LRT for multiplicative overdispersion manually;
DATA work.LRTb; * Binomial vs beta-binomial;
     DevBin=1531.7332; DevBB=534.1033;
     TestStat=DevBin-DevBB; TestDF=1;
     RegPvalue=1-PROBCHI(TestStat,TestDF);
     MixPvalue=(0.5*(1-PROBCHI(TestStat,TestDF)));
     FORMAT RegPValue MixPValue pvalue6.4; RUN;
TITLE1 "Likelihood Ratio Test for Binomial vs Beta-Binomial";
TITLE2 "Regular p-value uses DF, mixture p-value uses DF=DF,DF-1";
PROC PRINT NOOBS DATA=work.LRTb; RUN; TITLE1; TITLE2;
* Save corr of pred and actual propcorrect to square as R2;
TITLE "Correlation of Predicted and Actual PropCorrect";
PROC CORR NOSIMPLE DATA=work.BBpred OUT=work.Rpred;
     VAR PropCorrect; WITH yhat; RUN;
* Compute R2 in saved output;
DATA work.Rpred; SET work.Rpred;
     WHERE _TYPE_="CORR"; R2=PropCorrect*PropCorrect; RUN;
TITLE "R2 of Predicted and Actual PropCorrect";
PROC PRINT NOOBS DATA=work.Rpred; VAR R2; RUN; TITLE;
```

```
display "STATA Two-Predictor Beta-Binomial Model with Multiplicative Overdispersion"
display "Using betabin instead that has beta-binomial distribution"
betabin ncorrect c.nlivsli c.momed12, link(logit) n(ntrials)
display "-2LL=" e(ll)*-2  // Print -2LL for model
display "SAS scale factor= " 1/e(sigma) // Scale factor in SAS metric
estat ic, n(97)      // AIC and BIC matches SAS
display "LRT for multiplicative overdispersion (binomial vs beta-binomial)"
display "Test Statistic (df=1)= "           1531.7332-534.10333
display "Regular p-value= "      (1-chi2(1, 1531.7332-534.1033))
display "Mixture p-value= " 0.5*(1-chi2(1, 1531.7332-534.1033))
predict yhatBBpred  // Save predicted propcorrect per real person to dataset
corr yhatBBpred propcorrect // Get corr of pred and actual propcorrect
display "R2=" r(rho)^2      // Print R2 relative to empty model

display "STATA Two-Predictor Binomial Model with Multiplicative Overdispersion"
display "Using betabin and Getting Odds Ratios using eform"
betabin ncorrect c.nlivsli c.momed12, link(logit) n(ntrials) eform
```

```
print("R Two-Predictor Binomial Model using vglm with Multiplicative Overdispersion")
print("Parameter SEs do not match SAS and STATA exactly but are close")
ModelBetaBin = vglm(data=Example4a, betabinomial(lmu="logitlink", lrho="logitlink"),
                    formula=cbind(Ncorrect,Nincorrect)~1+NLIvSLI+momed12)
summary(ModelBetaBin)
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelBetaBin); AIC(ModelBetaBin); BIC(ModelBetaBin)

print("Likelihood Ratio Test for Addition of Multiplicative Overdispersion")
DevTestB=-2*(logLik(ModelBin)-logLik(ModelBetaBin))
RegPvalueB=pchisq((DevTestB), df=1, lower.tail=FALSE)
MixPvalueB=RegPvalueB/2
print("Test Statistic, Regular and Mixture P-values for DF=1")
DevTestB; RegPvalueB; MixPvalueB

print("Save predicted propcorrect and correlate with actual propcorrect")
Example4a$PredBetaBin = predict(ModelBetaBin, type="response")
rPredBetaBin = cor.test(Example4a$PredBetaBin, Example4a$PropCorrect, method="pearson")
print("R2"); rPredBetaBin$estimate^2
```

Partial SAS Output:

```
           Fit Statistics
-2 Log Likelihood             534.1
AIC  (Smaller is Better)      542.1
BIC  (Smaller is Better)      552.4
Pearson Statistic           71.6649 → when divided by DF=97, = 0.74, so pretty good!
```

```
         Parameter Estimates for Beta-Binomial Model
                               Standard
Effect            Estimate       Error    z Value    Pr > |z|
Intercept           2.9579      0.2500      11.83      <.0001 → proportion=.951
NLIvSLI            -0.9738      0.2729      -3.57      0.0004
momed12            0.04640     0.06855       0.68      0.4984
Scale Parameter     4.1434      0.9147                        → phi multiplier for variance (1=binomial?)

Likelihood Ratio Test for Binomial vs Beta-Binomial
Regular p-value uses DF, mixture p-value uses DF=DF,DF-1
                        Test     Test     Reg       Mix
 DevBin     DevBB       Stat      DF     Pvalue    Pvalue
1531.73    534.103    997.630      1     <.0001    <.0001

Pearson Correlation Coefficients, N = 97
       Prob > |r| under H0: Rho=0

                         Prop
                       Correct
yhat                   0.39662 → R^2 = 0.15731
Predicted Value         <.0001
```

Btw, PROC FMM has far fewer options for post-estimation (i.e., no CONTRAST or ESTIMATE). So I couldn't figure out how to get a multivariate Wald test for the two predictors jointly (i.e., for the model) ☹

Partial STATA Output:

```
Beta-binomial regression                          Number of obs   =         97
Link           = logit                            LR chi2(2)      =      13.61
Dispersion     = beta-binomial                    Prob > chi2     =     0.0035
Log likelihood = -267.05167                       Pseudo R2       =     0.0248
------------------------------------------------------------------------------
    ncorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nlivsli |  -.9737565   .2728606    -3.57   0.000    -1.508553   -.4389595
     momed12 |   .0464046   .0685461     0.68   0.498    -.0879434    .1807525
       _cons |   2.957862   .2500499    11.83   0.000     2.467773     3.44795 → proportion=.951
-------------+----------------------------------------------------------------
    /lnsigma |  -1.421521   .2207495    -6.44   0.000    -1.854182   -.9888596 = log(1/phi)
-------------+----------------------------------------------------------------
       sigma |   .2413467   .0532772                       .156581    .3720007 = 1/phi multiplier given by SAS
------------------------------------------------------------------------------
Likelihood-ratio test of sigma=0:  chibar2(01) =  997.63 Prob>=chibar2 = 0.000 → LRT for overdispersion

. display "SAS scale factor= " 1/e(sigma) // Scale factor in SAS metric
SAS scale factor= 4.1434165
. display "R2=" r(rho)^2      // Print R2 relative to empty model
R2=.15730629
LRT for multiplicative overdispersion (binomial vs beta-binomial)
Test Statistic (df=1)= 997.6299
Regular p-value= 0
Mixture p-value= 0
```

Partial R Output:

```
Coefficients:
               Estimate Std. Error z value              Pr(>|z|)
(Intercept):1  2.957854   0.255237 11.5887 < 0.00000000000000022  Beta0
(Intercept):2 -1.421511   0.214887 -6.6152      0.00000000003711  Log(1/phi) as given in Stata
NLIvSLI       -0.973743   0.271864 -3.5817             0.0003413
momed12        0.046404   0.071489  0.6491             0.5162707

Names of linear predictors: logitlink(mu), logitlink(rho) → not sure what rho is
Log-likelihood: -267.05167 on 190 degrees of freedom

[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelBetaBin)
[1] 534.10333
> AIC(ModelBetaBin)
[1] 542.10333
> BIC(ModelBetaBin)
[1] 552.40218
```

```
[1] "Likelihood Ratio Test for Addition of Multiplicative Overdispersion"
[1] "Test Statistic, Regular and Mixture P-values for DF=1"
> DevTestB
[1] 997.62984
> RegPvalueB
[1] 5.8810413e-219
> MixPvalueB
[1] 2.9405206e-219
[1] "Save predicted propcorrect and correlate with actual propcorrect"
[1] "R2"
0.15730603
```

5) Two-Predictor Binomial Model with Zero-Inflation (predicting number incorrect now)

Our negatively skewed data have one-inflation, not zero-inflation, but all the software routines I found were designed only for zero-inflation. So I solved this problem by predicting number incorrect instead of number correct. The model below says that number incorrect comes from a binomial distribution that has extra zero values. The "inflation" model that predicts the logit of being an "extra zero" is empty for now, because I just want to see how many extra zeros there are.

$$Logit\left(p_{ip} \text{ for incorrect trial}\right) = \beta_{0p} + \beta_{1p}(NLIvsSLI_i) + \beta_{2p}(MotherEd_i - 12)$$

$$Logit\left(p_{iz} \text{ for } y_i = 0\right) = \beta_{0z}$$

Conditional mean: $\#Incorrect_i \ = \left(Ntrials_i * p_{ip}\right) * p_{iz}$

I'm not even going to try to get the distributional notation or conditional variance right…

```
TITLE1 "SAS Two-Predictor Zero-Inflated Binomial Model";
TITLE2 "Using FMM and predicting Nincorrct instead";
PROC FMM DATA=work.Example4a NAMELEN=100;
     MODEL Nincorrect/Ntrials = NLIvSLI momed12 / LINK=LOGIT DIST=BINOMIAL;
     MODEL + / DIST=CONSTANT(0); * Empty inflation model predicting extra zero;
RUN; TITLE1; TITLE2;
* Compute LRT for zero-inflation manually;
DATA work.LRTc; * Binomial vs. zero-inflated binomial;
     DevBin=1531.7332; DevZBin=988.2183;
     TestStat=DevBin-DevZBin; TestDF=1;
     RegPvalue=1-PROBCHI(TestStat,TestDF);
     MixPvalue=(0.5*(1-PROBCHI(TestStat,TestDF)));
     FORMAT RegPValue MixPValue pvalue6.4; RUN;
TITLE1 "Likelihood Ratio Test for Binomial vs Zero-Inflated Binomial";
TITLE2 "Regular p-value uses DF, mixture p-value uses DF=DF,DF-1";
PROC PRINT NOOBS DATA=work.LRTc; RUN; TITLE1; TITLE2;

display "STATA Two-Predictor Zero-Inflated Binomial Model"
display "Use zbin and predict nincorrect instead"
display "ilink is for submodel predicting extra 0 (empty here)"
zib nincorrect c.nlivsli c.momed12, link(logit) n(ntrials) ilink(logit) inflate(_cons)
display "-2LL=" e(ll)*-2  // Print -2LL for model
estat ic, n(97)  // AIC and BIC matches SAS
display "LRT for zero-inflation (binomial vs zero-inflated binomial)"
display "Test Statistic (df=1)= "           1531.7332-988.2183
display "Regular p-value= "      (1-chi2(1, 1531.7332-988.2183))
display "Mixture p-value= " 0.5*(1-chi2(1, 1531.7332-988.2183))

print("R Two-Predictor Zero-Inflated Binomial Model using vglm Predicting Nincorrect")
print("Parameter SEs do not match SAS and STATA exactly but are close")
ModelZBin = vglm(data=Example4a, zibinomialff(lprob="logitlink", lonempstr0="logitlink",
                     multiple.responses=FALSE, ionempstr0=NULL, zero="onempstr0"),
                   formula=cbind(Nincorrect,Ncorrect)~1+NLIvSLI+momed12)
summary(ModelZBin)
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelZBin); AIC(ModelZBin); BIC(ModelZBin)

print("Likelihood Ratio Test for Addition of Zero-Inflation")
DevTestC=-2*(logLik(ModelBin)-logLik(ModelZBin))
RegPvalueC=pchisq((DevTestC), df=1, lower.tail=FALSE); MixPvalueC=RegPvalueC/2
print("Test Statistic, Regular and Mixture P-values for DF=1")
DevTestC; RegPvalueC; MixPvalueC
```

Partial SAS Output:

```
           Fit Statistics
-2 Log Likelihood              988.2 → -2LL diff = 543 relative to binomial, so zero-inflated is better
AIC  (Smaller is Better)       996.2
BIC  (Smaller is Better)      1006.5
Pearson Statistic              225.0 → Divided by DF=97, = 2.34375 (not as good)
Effective Parameters               4 → number of parameters in model
Effective Components               2 → This is a (confirmatory) mixture model

                 Parameter Estimates for Binomial Model
                                        Standard
   Component    Effect       Estimate       Error    z Value    Pr > |z|
           1    Intercept     -2.2099     0.08252     -26.78      <.0001  Beta0p
           1    NLIvSLI        0.6787     0.09347       7.26      <.0001  Beta1p
           1    momed12       -0.1149     0.02489      -4.61      <.0001  Beta2p

               Parameter Estimates for Mixing Probabilities
                            -----------------Linked Scale-----------------
              Mixing                          Standard
Component    Probability    Logit(Prob)         Error    z Value    Pr > |z|
        1         0.5878         0.3547        0.2063       1.72      0.0856
        2         0.4122        -0.3547                                        → Prob and logit of being an extra 0
```

Partial STATA Output:

```
Zero-inflated binomial regression                Number of obs    =         97
Regression link: logit                           Nonzero obs      =         57
Inflation link : logit                           Zero obs         =         40
                                                 LR chi2(2)       =     126.58
Log likelihood = -494.1091                       Prob > chi2      =     0.0000
------------------------------------------------------------------------------
  nincorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nincorrect   |
     nlivsli |   .6787023   .0934716     7.26   0.000     .4955014    .8619033  Beta1p
     momed12 |  -.1148639    .024894    -4.61   0.000    -.1636552   -.0660727  Beta2p
       _cons |  -2.209937   .0825224   -26.78   0.000    -2.371678   -2.048196  Beta0p
-------------+----------------------------------------------------------------
inflate      |                                                                 → logit of being extra 0
       _cons |  -.3547476   .2063317    -1.72   0.086    -.7591502     .049655  Beta0z
------------------------------------------------------------------------------
Test Statistic (df=1)= 543.5149
```

Partial R Output:

```
Coefficients:
               Estimate Std. Error  z value              Pr(>|z|)
(Intercept):1 -2.209921   0.068295 -32.3583 < 0.00000000000000022  Beta0p
(Intercept):2  0.354748   0.206329   1.7193               0.08555  Beta0z*-1 = logit of not extra 0
NLIvSLI        0.678734   0.084490   8.0333 0.0000000000000009488  Beta1p
momed12       -0.114872   0.022742  -5.0512 0.0000004391152171797  Beta2p

Names of linear predictors: logitlink(prob), logitlink(onempstr0)
Log-likelihood: -494.10915 on 190 degrees of freedom

[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelZBin)
[1] 988.2183
> AIC(ModelZBin)
[1] 996.2183
> BIC(ModelZBin)
[1] 1006.5171

[1] "Test Statistic, Regular and Mixture P-values for DF=1"
> DevTestC
[1] 543.51488
```

6) Two-Predictor **Beta-Binomial** Model with **Zero-Inflation** (predicting number incorrect now)

The model below says that number incorrect comes from a beta-binomial distribution that has extra zero values (instead of a binomial distribution that has extra zero values), allowing multiplicative overdispersion.

$$Logit\left(p_{ip}\text{ for incorrect}\right) = \beta_{0p} + \beta_{1p}(NLIvSLI_i) + \beta_{2p}(MotherEd_i - 12)$$
$$Logit\left(p_{iz}\text{ for }y_i = 0\right) = \beta_{0z}$$
$$\text{Conditional mean: }\#Incorrect_i = \left(Ntrials_i * p_{ip}\right) * p_{iz}$$

I'm not even going to try to get the distributional notation or conditional variance right…

```
TITLE1 "SAS Two-Predictor Zero-Inflated Beta-Binomial Model";
TITLE2 "Using FMM and predicting Nincorrect instead";
PROC FMM DATA=work.Example4a NAMELEN=100;
     MODEL Nincorrect/Ntrials = NLIvSLI momed12 / LINK=LOGIT DIST=BETABINOMIAL;
     MODEL + / DIST=CONSTANT(0); * Empty inflation model predicting extra zero;
     OUTPUT OUT=work.ZIBBpred PRED=yhat VAR=yhatvar;
RUN; TITLE1; TITLE2;
* Compute LRT for multiplicative overdispersion manually;
DATA work.LRTd; * Zero-inflated binomial vs beta-binomial;
     DevZBin=988.2183; DevZBB=527.5780;
     TestStat=DevZBin-DevZBB; TestDF=1;
     RegPvalue=1-PROBCHI(TestStat,TestDF);
     MixPvalue=(0.5*(1-PROBCHI(TestStat,TestDF)));
     FORMAT RegPValue MixPValue pvalue6.4; RUN;
TITLE1 "Likelihood Ratio Test for Zero-Inflated Binomial vs Beta-Binomial";
TITLE2 "Regular p-value uses DF, mixture p-value uses DF=DF,DF-1";
PROC PRINT NOOBS DATA=work.LRTd; RUN; TITLE1; TITLE2;
* Save corr of pred and actual propcorrect to square as R2;
TITLE "Correlation of Predicted and Actual PropIncorrect";
PROC CORR NOSIMPLE DATA=work.ZIBBpred OUT=work.Rpred;
     VAR PropIncorrect; WITH yhat; RUN;
* Compute R2 in saved output;
DATA work.Rpred; SET work.Rpred;
     WHERE _TYPE_="CORR"; R2=PropIncorrect*PropIncorrect; RUN;
TITLE "R2 of Predicted and Actual PropCorrect";
PROC PRINT NOOBS DATA=work.Rpred; VAR R2; RUN; TITLE;
```

```
display "STATA Two-Predictor Zero-Inflated Beta-Binomial Model"
display "Use zibbin and predict nincorrect instead"
zibbin nincorrect c.nlivsli c.momed12, link(logit) n(ntrials) ilink(logit) inflate(_cons)
display "-2LL=" e(ll)*-2  // Print -2LL for model
display "SAS scale factor= " 1/.1539883 // Scale factor in SAS metric
estat ic, n(97)  // AIC and BIC matches SAS
display "LRT for overdispersion (zero-inflated: binomial vs beta-binomial)"
display "Test Statistic (df=1)= "          988.2183-527.5780
display "Regular p-value= "     (1-chi2(1, 988.2183-527.5780))
display "Mixture p-value= " 0.5*(1-chi2(1, 988.2183-527.5780))
// Save predicted propincorrect per real person to dataset
predict yhatZIBB, xb  // Predicted outcome in logits
gen Npred=1/(1+exp(-1*yhatZIBB)) // Convert to probability
corr Npred propincorrect // Get corr of pred and actual propincorrect
display "R2=" r(rho)^2  // Print R2 relative to empty model
```

```
# Could not find zero-inflated beta-binomial regression in R, so I give up
```

Partial SAS Output:

```
           Fit Statistics
-2 Log Likelihood            527.6 → -2LL diff = 461 relative to ZI binomial, so ZI beta-binomial is better
AIC  (Smaller is Better)     537.6
BIC  (Smaller is Better)     550.5
Pearson Statistic          84.2707 → Divided by DF=97, = 0.869 (better)
Effective Parameters             5 → number of parameters in model
Effective Components             2 → still a (confirmatory) mixture model
```

```
                Parameter Estimates for Beta-Binomial Model
                                          Standard
   Component    Effect            Estimate       Error    z Value     Pr > |z|
           1    Intercept          -2.7505      0.3270      -8.41       <.0001 beta0p
           1    NLIvSLI             1.1282      0.3465       3.26       0.0011 beta1p
           1    momed12           -0.01789     0.08941      -0.20       0.8414 beta2p
           1    Scale Parameter     6.4940      1.6203                         → phi multiplier is bigger now

             Parameter Estimates for Mixing Probabilities
                          -----------------Linked Scale-----------------
                Mixing                      Standard
Component    Probability     Logit(Prob)        Error    z Value     Pr > |z|
        1         0.7494          1.0954       0.4370       2.51       0.0122
        2         0.2506         -1.0954                                       → Prob and Logit of being an extra 0

R2 of Predicted and Actual PropCorrect
   R2
0.14449
```

Partial STATA Output:

```
Zero-inflated beta-binomial regression             Number of obs     =         97
Regression link: logit                             Nonzero obs       =         57
Inflation link : logit                             Zero obs          =         40
                                                   LR chi2(2)        =      11.61
Log likelihood =  -263.789                         Prob > chi2       =     0.0030
------------------------------------------------------------------------------
  nincorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nincorrect   |
    nlivssli |   1.128224   .3464563     3.26   0.001     .4491825    1.807266 Beta1p
     momed12 |  -.0178967   .0894132    -0.20   0.841    -.1931434    .1573499 Beta2p
       _cons |  -2.750534   .3270209    -8.41   0.000    -3.391483   -2.109585 Beta0p
-------------+----------------------------------------------------------------
inflate      |                                                                 → logit of being an extra 0
       _cons |  -1.095397   .4369649    -2.51   0.012    -1.951832   -.2389614 Beta0z
-------------+----------------------------------------------------------------
    /lnsigma |  -1.870879   .2495082                     -2.359906   -1.381852 → log(1/phi)
-------------+----------------------------------------------------------------
       sigma |   .1539883   .0384213                      .0944291    .2511131 → 1/phi multiplier in SAS
------------------------------------------------------------------------------
Test Statistic (df=1)= 460.6403
```

---

7) <u>Four-Predictor</u> Beta-Binomial Model with Zero-Inflation (now predictors in inflation model)

This model adds our two predictors to the zero-inflation model (customizing the probability of being an extra zero).

$$Logit(p_i \text{ for incorrect}) = \beta_{0p} + \beta_{1p}(NLIvsSLI_i) + \beta_{2p}(MotherEd_i - 12)$$
$$Logit(p_{iz} \text{ for } y_i > 0) = \beta_{0z} + \beta_{1z}(NLIvsSLI_i) + \beta_{2z}(MotherEd_i - 12)$$
$$\text{Conditional mean: } \#Incorrect_i = (Ntrials_i * p_i) * p_{iz}$$

I'm not even going to try to get the distributional notation or conditional variance right…

```
display "STATA Two-Predictor Zero-Inflated Beta-Binomial Model"
display "Switch to zibbin and predict nincorrect instead"
display "Add two predictors of being extra zero"
zibbin nincorrect c.nlivsli c.momed12, link(logit) n(ntrials) ///
                    ilink(logit) inflate(c.nlivsli c.momed12)
display "-2LL=" e(ll)*-2  // Print -2LL for model
estat ic, n(97) // AIC and BIC matches SAS
```

Partial STATA Output only (SAS PROC FMM wouldn’t let me add zero-model predictors):

```
Zero-inflated beta-binomial regression            Number of obs    =         97
Regression link: logit                            Nonzero obs      =         57
Inflation link : logit                            Zero obs         =         40
                                                  LR chi2(2)       =       7.38
Log likelihood = -261.8274                        Prob > chi2      =     0.0249
------------------------------------------------------------------------------
  nincorrect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nincorrect   |
    nlivssli |   .3036772   .3546852     0.86   0.392     -.391493    .9988474  Beta1p
     momed12 |  -.2189386   .0812336    -2.70   0.007    -.3781535   -.0597237  Beta2p
       _cons |  -2.173967   .3963158    -5.49   0.000    -2.950731   -1.397202  Beta0p
-------------+----------------------------------------------------------------
inflate      |                                                                 → predict logit of extra 0
    nlivssli |  -3.970179   5.512301    -0.72   0.471    -14.77409    6.833733  Beta2z
     momed12 |  -.9569979   1.428802    -0.67   0.503    -3.757398    1.843402  Beta1z
       _cons |   .0198758   .6209887     0.03   0.974     -1.19724    1.236991  Beta0z
-------------+----------------------------------------------------------------
    /lnsigma |  -1.652934   .3139631                      -2.26829   -1.037578 → log(1/phi)
-------------+----------------------------------------------------------------
       sigma |   .1914873   .0601199                       .103489     .354312 → 1/phi multiplier in SAS
------------------------------------------------------------------------------
```

So which one should be pick? Let's do some informal model comparisons using distribution fit and relative fit (*may not be exactly comparable due to differences in estimation technique, but they should be close)

Two-Predictor Model	Pearson Chi-Square / DF	−2LL*	AIC*	BIC*
2. Regular Binomial	14.94	1531.73	1537.73	1545.46
3. +Additive Overdispersion	0.15	549.76	557.76	568.10
4. Beta-Binomial	0.74	534.10	542.10	552.40
5. Zero-Inflated Binomial	2.34	988.22	996.22	1006.52
6. Zero-Inflated Beta-Binomial	**0.87**	**527.58**	**537.58**	**550.45**
7. ZIBB + Predictors	?	523.65	537.65	555.68

Sample results using both programs (final model = zero-inflated beta-binomial without inflation predictors):

The extent that grammatical understanding (measured either as percent correct or percent incorrect; see below) at third grade could be predicted by language impairment group (non-specific=0, specific=1) and mother's years of education (centered such that 0=12 years) was examined in a series of generalized linear models. In the sample of N = 97 children, the mean percent correct was 0.92, with a large percentage of observations at or near the ceiling (1.00). Accordingly, we predicted the number of correct trials out of the number of possible trials using a logit link function to keep the predicted proportion correct outcomes bounded at 1. The type of model specifies that the number of correct responses follows a binomial-based distribution with 100 total trials and the model predicts the logit (log-odds) of a correct answer for any trial. Predicted outcomes in a logit metric can be translated into proportion correct via an inverse link function (which provides model-predicted proportions and their standard errors). All models were estimated using maximum likelihood within SAS GLIMMIX and FMM to assess distribution fit (or Stata glm, betabin, zib, and zibbin, or R vglm and glmer); predictor fixed effects were tested univariately using z-distributions without denominator degrees of freedom. Effect sizes are provided below as odds ratios: the exponentiated logit coefficient in which values from 0 to 1 indicate negative associations, 1 indicates no association, and values above 1 indicate positive associations.

Before interpreting our results, we tested the fit of models with alternative binomial-based conditional outcome distributions (each with main effects of group and mother's education) by examining the Pearson χ^2/DF statistic (in which 1=good fit), as well as likelihood ratio tests (i.e., treating −2 times the difference in log-likelihood between nested models as a χ^2 statistic with degrees of freedom equal to the number of additional parameters). As expected given the negatively skewed observed distribution, a model specifying a standard binomial distribution for number correct did not fit well (Pearson χ^2/DF = 14.94). Two methods of allowing overdispersion were then examined. First, we allowed additive overdispersion via an observation-level random intercept, which significantly improved model fit, −2ΔLL(1) = 987.97, p < .0001, but created a tendency towards underdispersion (Pearson χ^2/DF = 0.15). Second, we allowed multiplicative overdispersion by using a beta-binomial distribution, which significantly improved model fit, −2ΔLL(1) = 997.63, p < .0001, and appeared to fit well (Pearson χ^2/DF = 0.74). We then examined the potential for one-inflation by predicting number *incorrect* instead so that zero-inflation models could be fitted. A model predicting number incorrect with a zero-inflated binomial distribution was examined but did not fit as well (Pearson χ^2/DF = 2.34), although using a zero-inflated beta-binomial distribution instead did result in good fit (Pearson χ^2/DF = 0.87), as well as the lowest AIC and BIC of all the models. We also examined group and mother's education as predictors of zero-inflation but neither was significant (with higher AIC and BIC values), and thus the empty (unconditional) zero-inflation model was retained.

The model results indicated that 25.06% of the sample were predicted to be an extra 0 (i.e., to be part of the zero-inflated component of the distribution for number incorrect). Otherwise, the predicted intercept for a child with non-specific language impairment whose mother had 12 years of education was a logit = −2.75, which translates into percent incorrect = 0.06. Children with specific language impairment were predicted to have significantly more incorrect responses (logit = 1.12, OR = 3.09), although no significant difference was found for mother's years of education (logit = −0.02, OR = 0.98). The scale parameter for multiplicative overdispersion was 6.494, which was significant, −2ΔLL(1) = 460.64, p < .0001, relative to the zero-inflated binomial alternative.