Example 2: General Linear Models with a Single Quantitative or Binary Predictor
(complete syntax, data, and output available for SAS, STATA, and R electronically)

The data for this example were selected from the 2012 General Social Survey dataset (and were also used for Example 1). The current example will use general linear models to predict a single quantitative outcome (annual income in 1000s) from a quantitative predictor (a linear slope for years of education) and from a binary predictor (marital status: 0=unmarried and 1=married). It will also introduce how to obtain linear combinations of fixed effects to create predicted outcomes using SAS ESTIMATE, STATA LINCOM, and R GLHT.

Importing and Preparing Data for Analysis

In SAS:

```
* Paste in the folder address where "GSS_Example.xlsx" is saved after = before ;
%LET filesave= \\Client\C:\Dropbox\22SP_PSQF6243\PSQF6243_Example2;

* IMPORT GSS_Example.xlsx data using filesave reference using exact file name;
* from the Excel workbook in DATAFILE= location from SHEET= ;
* New SAS file is in "work" library place with name "Example2";
* "GETNAMES" reads in the first row as variable names;
* DBMS=XLSX (can also use EXCEL or XLS for .xls files);
PROC IMPORT DATAFILE="&filesave.\GSS_Example.xlsx"
            OUT=work.Example2 DBMS=XLSX REPLACE;
     SHEET="GSS_Example";
     GETNAMES=YES;
RUN;
```

Note: All SAS commands and comments end in a semi-colon!

```
* DATA = create new dataset, SET = from OLD dataset;
* So DATA + SET means "save as itself" after these actions;
* All data transformations must happen inside a DATA+SET+RUN combo;
DATA work.Example2; SET work.Example2;
* Label variables and apply value formats for variables used below;
* LABEL name=   "name: Descriptive Variable Label";
  LABEL marry=  "marry: Marital Status ((1=unmarried, 2=married)"
        educ=   "educ: Years of Education"
        income= "income: Annual Income in 1000s";
* Select cases complete on variables of interest;
  IF NMISS(income,educ,marry)>0 THEN DELETE;
RUN;
```

In STATA:

```
// Paste in the folder address where "GSS_Example.xlsx" is saved between " "
cd "\\Client\C:\Dropbox\22SP_PSQF6243\PSQF6243_Example2"

// IMPORT GSS_Example.xlsx data from working directory and exact file name
// To change all variable names to lowercase, remove "case(preserve")
clear // Clear before means close any open data
import excel "GSS_Example.xlsx", case(preserve) firstrow clear
// Clear after means re-import if it already exists (if need to start over)

// Label variables and apply value formats for variables used below
// label variable name "name: Descriptive Variable Label"
label variable marry   "marry: Marital Status (1=unmarried, 2=married)"
label variable educ    "educ: Years of Education"
label variable income  "income: Annual Income in 1000s"

// Select cases complete on variables of interest
egen nmiss = rowmiss(income educ marry)
drop if nmiss>0
```

In R:

```
# Set working directory (to import and export files to)
# Paste in the folder address where "GSS_Example.xlsx" is saved in quotes
setwd("C:/Dropbox/22SP_PSQF6243/PSQF6243_Example2")

# Import GSS_Example.xlsx data from working directory -- path = file name
Example2 = read_excel(path="GSS_Example.xlsx", sheet="GSS_Example")
# Convert to data frame to use for analysis
Example2 = as.data.frame(Example2)
# Labels added only as comments in R syntax file
```

Syntax for Descriptive Statistics and SAS Output:

```
TITLE "SAS Descriptive Statistics for Quantitative or Binary Variables";
PROC MEANS NDEC=3 N MEAN STDDEV VAR MIN MAX DATA=work.Example2;
     VAR income educ marry;
RUN; TITLE;
```

Because I added "VAR" to the list of statistics, I had to write all of them for SAS PROC MEANS.

```
display "STATA Descriptive Statistics for Quantitative or Binary Variables"
summarize income educ marry, detail

# describe prints sample descriptive statistics for quantitative variables
# list variables to be included in separate quotes within c concatenate function
print("R Descriptive Statistics for Quantitative for Quantitative or Binary Variables")
describe(x=Example2[ , c("income","educ","marry")])

# Get variances too (on diagonal of output matrix)
var(x=Example2[ , c("income","educ","marry")])
```

Variable	Label	N	Mean	Std Dev	Variance	Minimum	Maximum
income	income: Annual Income in 1000s	734	17.303	13.792	190.209	0.245	68.600
educ	educ: Years of Education	734	13.812	2.909	8.464	2.000	20.000
marry	marry: Marital Status (1=unmarried, 2=married)	734	1.459	0.499	0.249	1.000	2.000

Empty General Linear Model (no predictors):

$$Income_i = \beta_0 + e_i$$

NAMELEN extends printing of variable names; MODEL y = x / options (no x predictors so far); SOLUTION requests fixed effect solution be printed (oddly not a default), CLPARM provides confidence intervals (at alpha level), SS3 asks for Type 3 sums of squares only (not yet relevant)

To close the GLM procedure, you need both RUN; and QUIT; (seems redundant, but isn't)

In SAS:

```
TITLE "SAS Empty GLM Predicting Income";
PROC GLM DATA=work.Example2 NAMELEN=100;
     MODEL income = / SOLUTION ALPHA=.05 CLPARM SS3;
RUN; QUIT; TITLE;
```

Source	DF	Sum of Squares	**Mean Square**	F Value	Pr > F
Model	1	219751.8721	219751.8721	1155.32	<.0001
Error	733	139423.2319	**190.2090**		
Uncorrected Total	734	359175.1040			

Mean Square Error is the residual variance = 190.21 here. Stay tuned for what the rest means! 😊

R-Square	Coeff Var	Root MSE	income Mean
0.000000	79.70716	13.79163	17.30287

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	**17.30287466**	0.50905834	33.99	<.0001	16.30348846	18.30226086	**Beta0**

In STATA:

```
display "STATA GLM Empty Model Predicting Income"
regress income , level(95) // level gives (95)% CI for unstandardized solution
```

STATA's **regress** is general GLM routine. The first word after regress is the outcome variable. Level(95) requests 95% confidence intervals (the default). Below, MS stands for Mean Square (as in SAS above).

```
      Source |       SS           df       MS      Number of obs   =       734
-------------+----------------------------------   F(0, 733)       =      0.00
       Model |           0         0           .   Prob > F        =         .
    Residual |  139423.232       733  190.209048   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =    0.0000
       Total |  139423.232       733  190.209048   Root MSE        =    13.792

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   17.30287   .5090583    33.99   0.000     16.30349    18.30226 Beta0
------------------------------------------------------------------------------
```

In R:

```
print("R Empty GLM Predicting Income -- save as ModelEmpty")
ModelEmpty = lm(data=Example2, formula=income~1) # 1 represents intercept
anova(ModelEmpty)    # anova to print residual variance
summary(ModelEmpty) # summary to print fixed effects solution
confint.lm(ModelEmpty, level=.95) # confint to print level% CI for fixed effects
```

```
Analysis of Variance Table Response: income
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 733 139423 190.209
```

Mean Sq (Square) for "Residuals" = Residual Variance

```
Call: lm(formula = income ~ 1, data = Example2)
Coefficients:
            Estimate Std. Error t value              Pr(>|t|)
(Intercept) 17.30287    0.50906   33.99 < 0.0000000000000022 *** Beta0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 13.792 on 733 degrees of freedom

                2.5 %    97.5 %
(Intercept) 16.303488 18.302261
```

The output for an empty model differs slightly across programs. SAS counts the fixed intercept as part of the model sums of squares, whereas STATA and R do not, but they otherwise provide the same information.

In addition, STATA refers to the fixed intercept as **_cons**, which stands for constant. In models with more than one fixed effect, STATA will always list the fixed intercept LAST (much to my dismay).

Add a linear slope for a quantitative years of education predictor:
$$\boldsymbol{Income_i = \beta_0 + \beta_1 (Educ_i) + e_i}$$

In SAS:

```
TITLE "SAS GLM Predicting Income from Original Education";
PROC GLM DATA=work.Example2 NAMELEN=100;
     MODEL income = educ / SOLUTION ALPHA=.05 CLPARM SS3;
RUN; QUIT; TITLE;
```

```
                                        Sum of
Source                      DF         Squares    Mean Square    F Value    Pr > F
Model                        1      20634.9817     20634.9817     127.16    <.0001
Error                      732     118788.2502       162.2790
Corrected Total            733     139423.2319

R-Square     Coeff Var      Root MSE    income Mean
0.148002      73.62290      12.73888       17.30287

                                      Standard
Parameter          Estimate              Error    t Value    Pr > |t|      95% Confidence Limits
Intercept      -7.886678831         2.28277764      -3.45      0.0006     -12.36825087  -3.405106788 Beta0
educ            1.823745538         0.16173105      11.28      <.0001      1.506233517   2.141257559 Beta1
```

SAS no longer counts the fixed intercept as part of the model once 1+ predictors are added, so the SAS results will exactly match STATA and R. **Mean Square Error**, the residual variance, has been reduced to 162.28 after including education.

Interpret β_0 = intercept:

Interpret β_1 = slope of education:

In STATA:

```
display "STATA GLM Predicting Income from Original Education"
regress income educ, level(95)
```

```
      Source |       SS           df       MS      Number of obs   =       734
-------------+----------------------------------   F(1, 732)       =    127.16
       Model |  20634.9817         1  20634.9817   Prob > F        =    0.0000
    Residual |   118788.25       732   162.27903   R-squared       =    0.1480
-------------+----------------------------------   Adj R-squared   =    0.1468
       Total |  139423.232       733  190.209048   Root MSE        =    12.739

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.823746    .161731    11.28   0.000     1.506234    2.141258 Beta1
       _cons |  -7.886679   2.282778    -3.45   0.001    -12.36825   -3.405107 Beta0
------------------------------------------------------------------------------
```

STATA lists the fixed intercept last!

In R:

```
print("R GLM Predicting Income from Original Education -- save as ModelEduc")
ModelEduc = lm(data=Example2, formula=income~1+educ)
anova(ModelEduc)   # anova to print residual variance
summary(ModelEduc) # summary to print fixed effects solution
confint.lm(ModelEduc, level=.95) # confint.lm to print level% CI for fixed effects
```

```
Analysis of Variance Table  Response: income
           Df  Sum Sq  Mean Sq F value                 Pr(>F)
educ        1  20635 20634.98 127.157 < 0.0000000000000000222 ***
Residuals 732 118788    162.28 → Mean Square Residual = Residual Variance
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Call: lm(formula = income ~ 1 + educ, data = Example2)
Coefficients:
            Estimate Std. Error t value              Pr(>|t|)
(Intercept) -7.88668    2.28278 -3.4549             0.0005823 *** Beta0
educ         1.82375    0.16173 11.2764 < 0.00000000000000022 *** Beta1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 12.739 on 732 degrees of freedom
Multiple R-squared:   0.148,     Adjusted R-squared:  0.14684
F-statistic: 127.16 on 1 and 732 DF,  p-value: < 0.000000000000000222
```

```
                   2.5 %      97.5 %
(Intercept) -12.3682509 -3.4051068
educ          1.5062335  2.1412576
```

Given that no one had education = 0 years, let's center the education predictor so 0 now indicates 12 years to create a more meaningful model intercept ("you are here" sign as the model reference point).

Add a linear slope of a CENTERED quantitative years of education predictor:

$$Income_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$$

In SAS:

```
* Center education predictor so that 0 is meaningful;
DATA work.Example2; SET work.Example2;
     educ12=educ-12;
     LABEL educ12= "educ12: Education (0=12 years)";
RUN;

TITLE "SAS GLM Predicting Income from Centered Education (0=12)";
PROC GLM DATA=work.Example2 NAMELEN=100;
     MODEL income = educ12 / SOLUTION ALHPA=.05 CLPARM SS3;
* In SAS ESTIMATEs below, words refer to the estimated beta fixed effect,
  and values are the multiplier for the requested predictor value;
     ESTIMATE "Pred Income for  8 years (educ12=-4)" intercept 1 educ12 -4;
     ESTIMATE "Pred Income for 12 years (educ12= 0)" intercept 1 educ12  0;
     ESTIMATE "Pred Income for 16 years (educ12= 4)" intercept 1 educ12  4;
     ESTIMATE "Pred Income for 20 years (educ12= 8)" intercept 1 educ12  8;
RUN; QUIT; TITLE;
```

ESTIMATES will be explained on the next page!

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	20634.9817	20634.9817	127.16	<.0001
Error	732	118788.2502	162.2790		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.148002	73.62290	12.73888	17.30287

Mean Square Error, the residual variance, is still 162.28 because centering does not change the strength of prediction (but it does change β_0).

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	**13.99826762**	0.55404853	25.27	<.0001	12.91055398	15.08598127	**Beta0 new at 12**
educ12	**1.82374554**	0.16173105	11.28	<.0001	1.50623352	2.14125756	**Beta1 is same**

Interpret β_0 = intercept:

Interpret β_1 = slope of (education−12):

In STATA:

```
// Center education predictor so that 0 is meaningful
gen educ12=educ-12
label variable educ12 "educ12: Education (0=12 years)"

display "STATA GLM Predicting Income from Centered Education (0=12)"
regress income educ12, level(95)  // with 95% CI for unstandardized solution
```

```
      Source |       SS           df       MS      Number of obs   =       734
-------------+----------------------------------   F(1, 732)       =    127.16
       Model |  20634.9817         1  20634.9817   Prob > F        =    0.0000
    Residual |   118788.25       732   162.27903   R-squared       =    0.1480
-------------+----------------------------------   Adj R-squared   =    0.1468
       Total |  139423.232       733  190.209048   Root MSE        =    12.739

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      educ12 |   1.823746    .161731    11.28   0.000     1.506234    2.141258 Beta1 is same
       _cons |   13.99827   .5540485    25.27   0.000     12.91055    15.08598 Beta0 new at 12
------------------------------------------------------------------------------
```

In R:

```
# Center education predictor so that 0 is meaningful
Example2$educ12 = Example2$educ-12

print("R GLM Predicting Income from Centered Education 0=12 -- save as ModelEduc12")
ModelEduc12 = lm(data=Example2, formula=income~1+educ12)
anova(ModelEduc12)    # anova to print residual variance
summary(ModelEduc12) # summary to print fixed effects solution
confint.lm(ModelEduc12, level=.95) # confint.lm to print level% CI for fixed effects

Analysis of Variance Table Response: income
           Df Sum Sq  Mean Sq F value                Pr(>F)
educ12      1  20635 20634.98 127.157 < 0.0000000000000000222 ***
Residuals 732 118788    162.28 → Mean Square Residual = Residual Variance
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call: lm(formula = income ~ 1 + educ12, data = Example2)
Coefficients:
            Estimate Std. Error t value              Pr(>|t|)
(Intercept) 13.99827    0.55405  25.265 < 0.0000000000000002 *** Beta0 new at 12
educ12       1.82375    0.16173  11.276 < 0.0000000000000002 *** Beta1 is same
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 12.739 on 732 degrees of freedom
Multiple R-squared:   0.148,     Adjusted R-squared:  0.14684
F-statistic: 127.16 on 1 and 732 DF,  p-value: < 0.0000000000000002222

                 2.5 %     97.5 %
(Intercept) 12.9105540 15.0859813
educ12       1.5062335  2.1412576
```

The next set of commands in each program illustrate how to compute predicted $\hat{y}_i$ outcomes given any value(s) of the predictor(s). Model: $Income_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$

Predicted income for 8 years education: $\hat{y}_i = 14.00(1) + 1.82(-4) = 6.70$
Predicted income for 12 years education: $\hat{y}_i = 14.00(1) + 1.82(0) = 14.00$
Predicted income for 16 years education: $\hat{y}_i = 14.00(1) + 1.82(4) = 21.29$
Predicted income for 20 years education: $\hat{y}_i = 14.00(1) + 1.82(8) = 28.59$

```
* In SAS ESTIMATEs below, words refer to the estimated beta fixed effect,
  and values are the multiplier for the requested predictor value;
    ESTIMATE "Pred Income  8 years (educ12=-4)" intercept 1 educ12 -4;
    ESTIMATE "Pred Income 12 years (educ12= 0)" intercept 1 educ12  0;
    ESTIMATE "Pred Income 16 years (educ12= 4)" intercept 1 educ12  4;
    ESTIMATE "Pred Income 20 years (educ12= 8)" intercept 1 educ12  8;
```

```
// In STATA LINCOMs below, _cons is intercept, words refer to the beta fixed effect,
// and values are the multiplier for the requested predictor value
lincom _cons*1 + educ12*-4  // Pred Income for  8 years (educ12=-4)
lincom _cons*1 + educ12*0   // Pred Income for 12 years (educ12= 0)
lincom _cons*1 + educ12*4   // Pred Income for 16 years (educ12= 4)
lincom _cons*1 + educ12*8   // Pred Income for 18 years (educ12= 8)
```

```
print("R Demonstrating how to get predicted outcomes using glht -- save as PredEduc12")
print("In number lists below, values are multiplier for each fixed effect in order")
PredEduc12 = glht(model=ModelEduc12, linfct=rbind(
  "Pred Income at  8 years (educ12=-4)" = c(1,-4),
  "Pred Income at 12 years (educ12= 0)" = c(1, 0),
  "Pred Income at 16 years (educ12= 4)" = c(1, 4),
  "Pred Income at 20 years (educ12= 8)" = c(1, 8)))
print("Print glht linear combination results with unadjusted p-values")
summary(PredEduc12, test=adjusted("none"))
confint(PredEduc12, level=.95, calpha=univariate_calpha())
```

These are the results from SAS ESTIMATEs:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Pred Income for 8 years (educ12=-4)	**6.7032855**	1.05102297	6.38	<.0001	4.6399066	8.7666643
Pred Income for 12 years (educ12= 0)	**13.9982676**	0.55404853	25.27	<.0001	12.9105540	15.0859813
Pred Income for 16 years (educ12= 4)	**21.2932498**	0.58848286	36.18	<.0001	20.1379343	22.4485652
Pred Income for 20 years (educ12= 8)	**28.5882319**	1.10574690	25.85	<.0001	26.4174185	30.7590454

These are the results from STATA LINCOMs:

```
. lincom _cons*1 + educ12*-4  // Pred Income for  8 years (educ12=-4)
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   6.703285   1.051023     6.38   0.000     4.639907    8.766664
------------------------------------------------------------------------------

. lincom _cons*1 + educ12*0   // Pred Income for 12 years (educ12= 0)
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   13.99827   .5540485    25.27   0.000     12.91055    15.08598
------------------------------------------------------------------------------

. lincom _cons*1 + educ12*4   // Pred Income for 16 years (educ12= 4)
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   21.29325   .5884829    36.18   0.000     20.13793    22.44857
------------------------------------------------------------------------------

. lincom _cons*1 + educ12*8   // Pred Income for 18 years (educ12= 8)
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   28.58823   1.105747    25.85   0.000     26.41742    30.75905
------------------------------------------------------------------------------
```

These are the results from R GLHTs:

```
Linear Hypotheses:
                                         Estimate Std. Error t value                Pr(>|t|)
Pred Income for  8 years (educ12=-4) == 0  6.70329    1.05102  6.3779       0.0000000003181 ***
Pred Income for 12 years (educ12= 0) == 0 13.99827    0.55405 25.2654 < 0.00000000000000022 ***
Pred Income for 16 years (educ12= 4) == 0 21.29325    0.58848 36.1833 < 0.00000000000000022 ***
Pred Income for 20 years (educ12= 8) == 0 28.58823    1.10575 25.8542 < 0.00000000000000022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- none method)
        Simultaneous Confidence Intervals
```

```
                                             Estimate lwr      upr
Pred Income at  8 years (educ12=-4) == 0  6.70329  4.63991  8.76666
Pred Income at 12 years (educ12= 0) == 0 13.99827 12.91055 15.08598
Pred Income at 16 years (educ12= 4) == 0 21.29325 20.13793 22.44857
Pred Income at 20 years (educ12= 8) == 0 28.58823 26.41742 30.75905
```

Standardized Solution for Education Predicting Income: Results using standardized variables (z-scored income and education), in which fixed slopes are in a correlation metric (−1 to 1)

In SAS:

```
TITLE1 "SAS GLM Predicting Income from Centered Education";
TITLE2 "Using REG instead of GLM to get standardized Effects";
PROC REG DATA=work.Example2;
     MODEL income = educ12 / STB; * STB gives standardized solution;
RUN; QUIT; TITLE1; TITLE2;
```

```
                                    Parameter Estimates
                                        Parameter     Standard                          Standardized
Variable   Label                   DF    Estimate        Error   t Value   Pr > |t|         Estimate
Intercept  Intercept                1    13.99827      0.55405     25.27     <.0001                0 Beta0
educ12     Education (0=12 years)   1     1.82375      0.16173     11.28     <.0001          0.38471 Beta1
```

In STATA:

```
display "STATA GLM Predicting Income from Centered Education (0=12)"
regress income educ12, beta  // beta gives standardized solution
```

```
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      educ12 |   1.823746    .161731    11.28   0.000                 .3847109 Beta1
       _cons |   13.99827   .5540485    25.27   0.000                        . Beta0 (=0)
------------------------------------------------------------------------------
```

In R:

```
print("R GLM Predicting Income using Standardized Solution -- save as ModelEducSTD")
print("scale () standardizes each variable as M=0 SD=1 z-score for analysis")
ModelEducSTD = lm(data=Example2, formula=scale(income)~1+scale(educ12))
summary(ModelEducSTD) # print standardized fixed effect solution
```

```
Coefficients:
                             Estimate               Std. Error t value             Pr(>|t|)
(Intercept)    -0.00000000000000011221  0.03409318589249633186   0.000                    1      Beta0
scale(educ12)   0.38471088238443779117  0.03411643389103318630  11.276 <0.0000000000000002 *** Beta1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add a linear slope for dummy-coded marital status predictor:

$$Income_i = \beta_0 + \beta_1(Marry01_i) + e_i$$

<u>Results will be:</u>

Predicted income unmarried (marry01=0): $\hat{y}_i = 14.45(1) + 6.22(0) = 14.45$

Predicted income unmarried (marry01=1): $\hat{y}_i = 14.45(1) + 6.22(1) = 20.67$

In SAS:

```
* Recode marry predictor so that 0 is meaningful;
DATA work.Example2; SET work.Example2;
     marry01=.; * Create new empty variable, then recode;
     IF marry=1 THEN marry01=0;
     IF marry=2 THEN marry01=1;
     LABEL marry01= "marry01: 0=unmarried, 1=married";
RUN;

TITLE "SAS GLM Predicting Income from Marry01 (0=Unmarried,1=Married)";
PROC GLM DATA=work.Example2 NAMELEN=100;
     MODEL income = marry01 / SOLUTION ALPHA=.05 CLPARM SS3;
* ESTIMATEs below request predicted outcome means for each group;
     ESTIMATE "Pred Income for Unmarried (marry01=0)" intercept 1 marry01 0; * Beta0;
     ESTIMATE "Pred Income for Married   (marry01=1)" intercept 1 marry01 1; * Beta0+Beta1;
RUN; QUIT; TITLE;
```

Source	DF	Sum of Squares	**Mean Square**	F Value	Pr > F
Model	1	7060.1016	7060.1016	39.04	<.0001
Error	732	132363.1303	**180.8239**		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.050638	77.71587	13.44708	17.30287

Mean Square Error, the residual variance, has been reduced to 180.82 after including education.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	**14.44543451**	0.67488958	21.40	<.0001	13.12048450	15.77038452	**Beta0**
marry01	**6.22362335**	0.99601482	6.25	<.0001	4.26823703	8.17900967	**Beta1**

These are the extra linear combinations of the fixed effects created by SAS ESTIMATEs:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Pred Income for Unmarried=0)	**14.4454345**	0.67488958	21.40	<.0001	13.1204845	15.7703845
Pred Income for Married=1	**20.6690579**	0.73250910	28.22	<.0001	19.2309886	22.1071271

Interpret β_0 = intercept:

Interpret β_1 = slope of marry01:

In STATA:

```
// Recode marry predictor so that 0 is meaningful
gen marry01=. // Create new empty variable, then recode
replace marry01=0 if marry==1
replace marry01=1 if marry==2
label variable marry01 "marry01: 0=unmarried, 1=married"

display "STATA GLM Predict Income from Marry01 (0=Unmarried,1=Married)"
regress income marry01, level(95) // with 95% CI for unstandardized solution
lincom _cons*1 + marry01*0 // Pred Income for Unmarried=0 = Beta0
lincom _cons*1 + marry01*1 // Pred Income for Married=1   = Beta0 + Beta1
```

Source	SS	df	**MS**
Model	7060.10161	1	7060.10161
Residual	132363.13	732	**180.823948**
Total	139423.232	733	190.209048

Number of obs	=	734
F(1, 732)	=	39.04
Prob > F	=	0.0000
R-squared	=	0.0506
Adj R-squared	=	0.0493
Root MSE	=	13.447

```
-----------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
     marry01 |   6.223623   .9960148     6.25   0.000     4.268237     8.17901 Beta1
       _cons |   14.44543   .6748896    21.40   0.000     13.12048    15.77038 Beta0
-----------------------------------------------------------------------------
```

These are the extra linear combinations of the fixed effects created by STATA LINCOMs:

```
. lincom _cons*1 + marry01*0 // Pred Income for Unmarried=0 = Beta0
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   14.44543   .6748896    21.40   0.000     13.12048    15.77038
------------------------------------------------------------------------------

. lincom _cons*1 + marry01*1 // Pred Income for Married=1 = Beta0 + Beta1
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   20.66906   .7325091    28.22   0.000     19.23099    22.10713
------------------------------------------------------------------------------
```

In R:

```
# Recode marry predictor so that 0 is meaningful
Example2$marry01=NA  # Create new empty variable, then recode
Example2$marry01[which(Example2$marry==1)]=0
Example2$marry01[which(Example2$marry==2)]=1

print("R GLM Predicting Income from Marry01 (0=Unmarried,1=Married) -- save ModelMarry01")
ModelMarry01 = lm(data=Example2, formula=income~1+marry01)
anova(ModelMarry01)    # anova to print residual variance
summary(ModelMarry01) # summary to print fixed effects solution
confint.lm(ModelMarry01, level=.95) # confint.lm to print level% CI for fixed effects

print("R Demonstrating how to get predicted outcomes using glht -- save as PredMarry01")
print("In number lists below, values are multiplier for each fixed effect in order")
PredMarry01 = glht(model=ModelMarry01, linfct=rbind(
  "Pred Income for Unmarried=0" = c(1,0),
  "Pred income for Married=1"   = c(1,1)))
print("Print glht linear combination results with unadjusted p-values")
summary(PredMarry01, test=adjusted("none"))
confint(PredMarry01, level=.95, calpha=univariate_calpha())
```

```
Analysis of Variance Table
Response: income
           Df   Sum Sq Mean Sq F value           Pr(>F)
marry01     1   7060.1 7060.10 39.0441 0.00000000070292 ***
Residuals 732 132363.1  180.82 → Mean Square Residual = Residual Variance
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call: lm(formula = income ~ 1 + marry01, data = Example2)
Coefficients:
            Estimate Std. Error t value              Pr(>|t|)
(Intercept) 14.44543    0.67489 21.4041 < 0.0000000000000022 ***
marry01      6.22362    0.99601  6.2485       0.0000000007029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 13.447 on 732 degrees of freedom
Multiple R-squared:  0.050638,    Adjusted R-squared:  0.049341
F-statistic: 39.044 on 1 and 732 DF,  p-value: 0.00000000070292

                2.5 %     97.5 %
(Intercept) 13.120484 15.7703845
marry01      4.268237  8.1790097
```

```
        Simultaneous Tests for General Linear Hypotheses
Linear Hypotheses:
                                 Estimate Std. Error t value              Pr(>|t|)
Pred Income for Unmarried=0 == 0 14.44543    0.67489  21.404 < 0.00000000000000022 ***
Pred income for Married=1 == 0   20.66906    0.73251  28.217 < 0.00000000000000022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- none method)

        Simultaneous Confidence Intervals
Linear Hypotheses:
                                 Estimate lwr      upr
Pred Income for Unmarried=0 == 0 14.44543 13.12048 15.77038
Pred income for Married=1 == 0   20.66906 19.23099 22.10713
```

One last thing: To get a Cohen's d effect size for the mean income difference between unmarried and married persons, we can calculate d from the t test-statistic: $d = \frac{2t}{\sqrt{DF_{den}}} = \frac{2*6.25}{\sqrt{732}} = 0.462$ → mean income is about 0.462 standard deviations higher for married than unmarried persons.

In SAS:

```
* Compute Cohen d effect size for marry01 from t test-statistic;
DATA work.MakeD;
     CohenD=2*6.25/SQRT(732);
RUN;
TITLE "SAS Print result for Cohen d";
PROC PRINT NOOBS DATA=work.MakeD;
RUN;
```

The code on the left makes a new dataset, creates a new variable d for the result of the formula, and then PROC PRINT outputs that new dataset.

```
CohenD
0.46201
```

In STATA:

```
display "STATA Compute Cohen's D from t test-statistic"
display 2*6.25/sqrt(732)
.46201329
```

In R:

```
print("R Compute d effect size for marry01 from t test-statistic")
CohenD = 2*6.25/sqrt(732)
print(CohenD)
[1] 0.46201329
```

Example Results Section:

The extent to which annual income in thousands of dollars (M = 17.30, SD = 13.79) could be predicted from years of education (M = 13.81, SD = 2.91) and binary marital status (1 = unmarried 54.09%, 2 = married 45.91%) was examined in separate general linear models (i.e., simple linear regressions).

To create a meaningful model intercept, education was centered such that 0 = 12 years. Education was found to be a significant predictor of annual income: Relative to the reference expected income for a person with 12 years of education provided by the model intercept of 14.00k (SE = 0.55), for every additional year of education, annual income was expected to be higher by 1.82k (SE = 0.16, p < .001), resulting in a standardized coefficient = 0.38 (i.e., the Pearson correlation between annual income and education). For example, persons with only 8 years of education were predicted to have an annual income of only 6.70k (SE = 1.05), persons with 16 years of education were predicted to have an annual income of 21.29k (SE = 0.59), and persons with 20 years of education were predicted to have an annual income of 28.59k (SE = 1.11). *[Spoiler alert: we will test the adequacy of only a linear (constant) effect for years of education in example 3.]*

We then examined prediction of annual income by binary marital status. To create a meaningful model intercept, marital status was dummy-coded so that 0 = unmarried persons and 1 = married persons. Marital status was also a significant predictor of annual income: Relative to the reference expected income for unmarried persons provided by the model intercept of 14.45k (SE = 0.67), married persons were expected to have significantly greater income by 6.22k (SE = 1.00, $p < .001$), resulting in a predicted income for married persons of 20.67k (SE = 0.73) and a standardized mean difference of Cohen's $d = 0.462$.

Note: because a GLM with a single binary predictor is also known as a "two-sample t-test" here is what the results would look like written from that angle… A two-sample *t*-test (i.e., assuming homogeneous variance across groups) was used to examine mean differences between unmarried and married persons in annual income. A significant mean difference was found, $t(732) = 6.25$, $p < .001$, such that annual income for married persons ($M = 20.67$k, SE = 0.73) was significantly higher than for unmarried persons ($M = 14.45$k, SE = 0.67).

Bonus: Bivariate Pearson Correlation Matrix, Significance Tests, and Confidence Intervals

In SAS:

```
TITLE "SAS Pearson Correlations and CIs";
PROC CORR NOSIMPLE DATA=work.Example2 PEARSON FISHER(BIASADJ=NO ALPHA=.05);
     VAR income educ marry;
RUN; TITLE;
```

```
                Pearson Correlation Coefficients, N = 734
                        Prob > |r| under H0: Rho=0

                                       income        educ        marry
income                                1.00000     0.38471      0.22503
income: Annual Income in 1000s                     <.0001       <.0001
educ                                  0.38471     1.00000      0.05112
educ: Years of Education               <.0001                   0.1665
marry                                 0.22503     0.05112      1.00000
marry: 2-Category Marital Status       <.0001      0.1665
```

```
                      Pearson Correlation Statistics (Fisher's z Transformation)
          With                Sample                                           p Value for
Variable  Variable    N   Correlation  Fisher's z    95% Confidence Limits       H0:Rho=0
income    educ      734       0.38471     0.40558     0.321290     0.444696       <.0001
income    marry     734       0.22503     0.22895     0.155191     0.292629       <.0001
educ      marry     734       0.05112     0.05116    -0.021326     0.123028       0.1666
```

In STATA:

```
display "STATA Pearson Correlations and CIs"
pwcorr income educ marry, sig
```

```
             |   income     educ    marry
-------------+---------------------------
      income |   1.0000
             |
        educ |   0.3847   1.0000
             |   0.0000
       marry |   0.2250   0.0511   1.0000
             |   0.0000   0.1665
```

```
// To get CI using r-to-z, need to download and run a special module
ssc install ci2
ci2 income educ, corr
ci2 income marry, corr
ci2 educ marry, corr
```

```
ci2 income educ, corr

Confidence interval for Pearson's product-moment correlation of income and educ, based on Fisher's
transformation. Correlation = 0.385 on 734 observations (95% CI: 0.321 to 0.445)

. ci2 income marry, corr

Confidence interval for Pearson's product-moment correlation of income and marry, based on Fisher's
transformation. Correlation = 0.225 on 734 observations (95% CI: 0.155 to 0.293)

. ci2 educ marry, corr

Confidence interval for Pearson's product-moment correlation of educ and marry, based on Fisher's
transformation. Correlation = 0.051 on 734 observations (95% CI: -0.021 to 0.123)
```

In R:

```
print("R Pearson Correlation Matrix")
cor(x=cbind(Example2$income,Example2$educ,Example2$marry), method="pearson")
```

```
           [,1]        [,2]        [,3]
[1,] 1.00000000 0.384710882 0.225028696
[2,] 0.38471088 1.000000000 0.051118354
[3,] 0.22502870 0.051118354 1.000000000
```

```
print("R Pearson Correlation Pairwise Significance tests and CIs")
cor.test(x=Example2$income, y=Example2$educ,  method="pearson", conf.level=.95)
cor.test(x=Example2$income, y=Example2$marry, method="pearson", conf.level=.95)
cor.test(x=Example2$educ,   y=Example2$marry, method="pearson", conf.level=.95)
```

```
data:  Example2$income and Example2$educ
t = 11.2764, df = 732, p-value < 0.0000000000000000222
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.32129033 0.44469587
sample estimates:
       cor
0.38471088

data:  Example2$income and Example2$marry
t = 6.24852, df = 732, p-value = 0.00000000070292
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.15519069 0.29262863
sample estimates:
      cor
0.2250287

data:  Example2$educ and Example2$marry
t = 1.38484, df = 732, p-value = 0.16652
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: -0.021325704  0.123028418
sample estimates:
        cor
0.051118354
```