Example 3: General Linear Models with a Single Predictor in SAS and STATA

The data for this example were selected from the 2012 General Social Survey dataset featured in Mitchell (2015); these data were also used for examples 1 and 2. The current example will use general linear models to predict a single quantitative outcome (annual income in 1000s) from a quantitative predictor (a linear effect of years of education) and from a binary predictor (marital status: 0=unmarried and 1=married).

SAS Syntax for Importing and Preparing Data for Analysis:

```
* Paste in the folder address where "GSS_Example.xlsx" is saved after = before ;
%LET filesave= \\Client\C:\Dropbox\21SP_PSQF6242\PSQF6242_Example3;

* IMPORT GSS_Example.xlsx data using filesave reference and exact file name;
* from the Excel workbook in DATAFILE= location from SHEET= ;
* New SAS file is in "work" library place with name "Example3";
* "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.Example3 DBMS=XLSX REPLACE;
     SHEET="GSS_Example";
     GETNAMES=YES;
RUN;
* Create formats: set of value labels for categorical variables;
PROC FORMAT;
     VALUE Fmarry 1="1.Unmarried" 2="2.Married";
RUN;
* 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.Example3; SET work.Example3;
* Label variables and apply value formats for variables used below;
* LABEL name=    "name: Descriptive Variable Label";
  LABEL marry=   "marry: 2-Category Marital Status"
        educ=    "educ: Years of Education"
        income=  "income: Annual Income in 1000s";
* Apply value labels created above: name Format.;
  FORMAT marry Fmarry.;
* Select cases complete on variables of interest;
  IF NMISS(income,educ,marry)>0 THEN DELETE;
RUN;
```

All SAS commands and comments end in a semi-colon.

STATA Syntax for Importing and Preparing Data for Analysis:

```
// Paste in the folder address where "GSS_Example.xlsx" is saved between " "
   global filesave "C:\Dropbox\21SP_PSQF6242\PSQF6242_Example3"

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

// Create formats: set of value labels for categorical variables;
   label define Fmarry    1 "1.Unmarried" 2 "2.Married"
// Label variables and apply value formats for variables used below
// label variable name    "name: Descriptive Variable Label"
   label variable marry   "marry: 2-Category Marital Status"
   label variable educ    "educ: Years of Education"
   label variable income  "income: Annual Income in 1000s"
// Apply value labels created above: name Format
   label values marry Fmarry
// Select cases complete on variables of interest
   egen nmiss = rowmiss(income educ marry)
   drop if nmiss>0
```

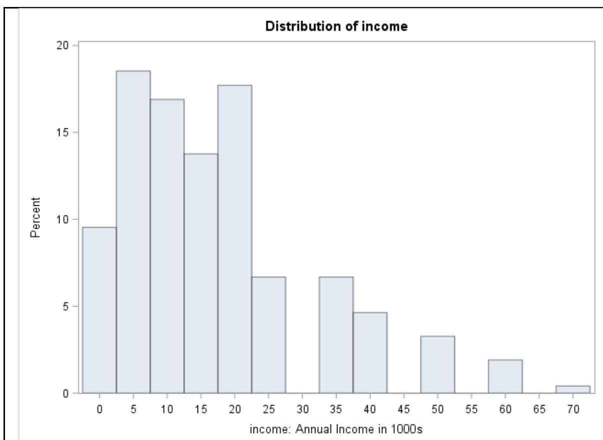
Syntax for Creating Descriptive Statistics, Histograms, and SAS Output:

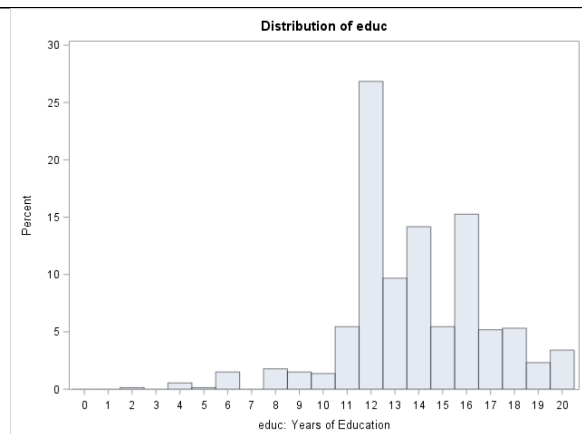
```
TITLE "SAS Descriptive Statistics for Quantitative Variables";
PROC MEANS NDEC=3 NOLABELS N MEAN STDDEV VAR MIN MAX DATA=work.Example3;
     VAR income educ;
RUN; TITLE;

display "STATA Descriptive Statistics for Quantitative Variables"
format income educ %5.3f  // format used to print three digits
summarize income educ, format detail
```

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

Variable	N	Mean	Std Dev	Variance	Minimum	Maximum
income	734	17.303	13.792	190.209	0.245	68.600
educ	734	13.812	2.909	8.464	2.000	20.000

```
* Histograms to visualize quantitative variables;
* NOPRINT spares the rest of the results I do not want right now;
TITLE "SAS Histograms of Quantitative Variables";
PROC UNIVARIATE NOPRINT DATA=work.Example3;
     VAR income educ;
     HISTOGRAM income / MIDPOINTS=0 TO 70 BY 5;
     HISTOGRAM educ   / MIDPOINTS=0 TO 20 BY 1;
RUN; QUIT; TITLE;

display "STATA Histograms of Quantitative Variables"
histogram income, percent discrete width(5) start(0)
histogram educ,   percent discrete width(1) start(0)
```






```
TITLE "SAS Descriptive Statistics for Categorical Variable";
PROC FREQ DATA=work.Example3;
     TABLE marry;
RUN; TITLE;

display "STATA Descriptive Statistics for Categorical Variable"
tabulate marry
```

marry: 2-Category Marital Status

marry	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.Unmarried	397	54.09	397	54.09
2.Married	337	45.91	734	100.00

Syntax and SAS Output for Pearson Correlation Matrix:

```
TITLE "SAS Pearson Correlations and CIs";
PROC CORR NOSIMPLE DATA=work.Example3 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)

Variable	With Variable	N	Sample Correlation	Fisher's z	95% Confidence Limits		p Value for 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

```
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)
```

Syntax and Selected Output for General Linear Models

Empty Model (no predictors): $Income_i = \beta_0 + e_i$

In SAS:

```
TITLE "SAS GLM Empty Model PredictingIncome";
PROC GLM DATA=work.Example3 NAMELEN=100;
     MODEL income = / SOLUTION ALPHA=.05 CLPARM;
RUN; QUIT; TITLE;
```

> NAMELEN extends printing of variable names; MODEL y = x / options (no x predictors so far); CLPARM provides confidence intervals (at chosen alpha level), SOLUTION requests fixed effect solution be printed (oddly not a default)
>
> To close the GLM procedure, you need both RUN; and QUIT; (seems redundant, but isn't)

```
The GLM Procedure
Dependent Variable: income   income: Annual Income in 1000s

                                          Sum of
Source                     DF            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

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

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

> **Mean Square Error** (Mean Square **Residual** in STATA) gives the residual variance = 190.21 here. We will discuss what the rest of this output means in GLMs with multiple predictors.

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
------------------------------------------------------------------------------
```

SAS and STATA's output for an empty model differ slightly: SAS counts the fixed intercept as part of the model sums of squares, whereas STATA does not… but they otherwise provide the same information.

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

Add a linear effect of a quantitative predictor for education: $Income_i = \beta_0 + \beta_1(Educ_i) + e_i$

In SAS:

```
TITLE "SAS GLM Predicting Income from Original Education";
PROC GLM DATA=work.Example3 NAMELEN=100;
     MODEL income = educ / SOLUTION ALPHA=.05 CLPARM;
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
```

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

```
                                     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
```

In STATA:

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

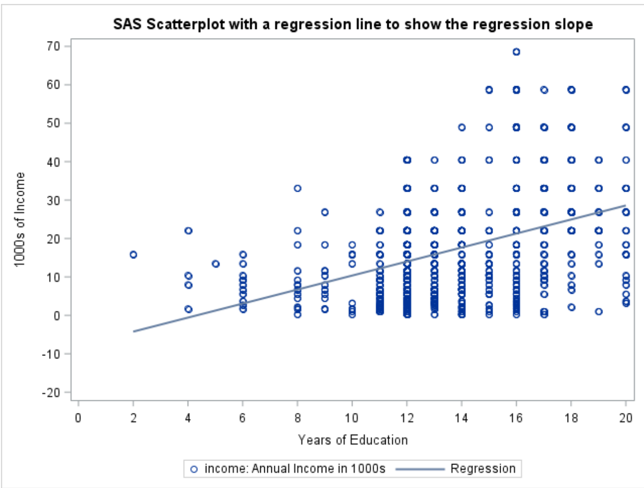
```
      Source |       SS           df       MS      Number of obs   =       734
-------------+----------------------------------   F(1, 732)       =    127.16
       Model |  20634.9817         1  20634.9817   Prob > F        =    0.0000
    Residual |   118788.25       732   162.27903   R-squared       =    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!

Interpret β_0 = intercept:

Interpret β_1 = slope of education:

```
TITLE "SAS Scatterplot with a regression
       line to show the regression slope";
PROC SGPLOT DATA=work.Example3;
     SCATTER x=educ y=income;
     REG     x=educ y= income;
     XAXIS LABEL="Years of Education"
           VALUES=(0 TO 20 BY 5);
     YAXIS LABEL="1000s of Income"
           VALUES=(-20 TO 70 BY 20);
RUN;
```



Given that no one had education = 0 years, let's replace the education predictor with a new centered version, in which 0 now indicates 12 years, to create a meaningful model intercept ("you are here" sign as the model reference point): $Income_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$

Using Centered Education Predictor in SAS:

```
* Center education predictor so that 0 is meaningful;
DATA work.Example3; SET work.Example3;
     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.Example3 NAMELEN=100 PLOTS(UNPACK)=DIAGNOSTICS;
     MODEL income = educ12 / SOLUTION ALHPA=.05 CLPARM;
* In 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;
RUN; QUIT; TITLE;
```

PLOTS option makes all kinds of figures for diagnosing model mis-specification (stay tuned).

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 beta0).

This is the regular table of fixed effects estimated directly by the model (WILL BE LAST):

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**

The ESTIMATE commands provide an example of how to compute predicted values for the outcome 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: $\hat{y}_i = 14.00 + 1.82(-4) = 6.70$
Predicted income for 12 years education: $\hat{y}_i = 14.00 + 1.82(0) = 14.00$
Predicted income for 16 years education: $\hat{y}_i = 14.00 + 1.82(4) = 21.29$
Predicted income for 20 years education: $\hat{y}_i = 14.00 + 1.82(8) = 28.59$

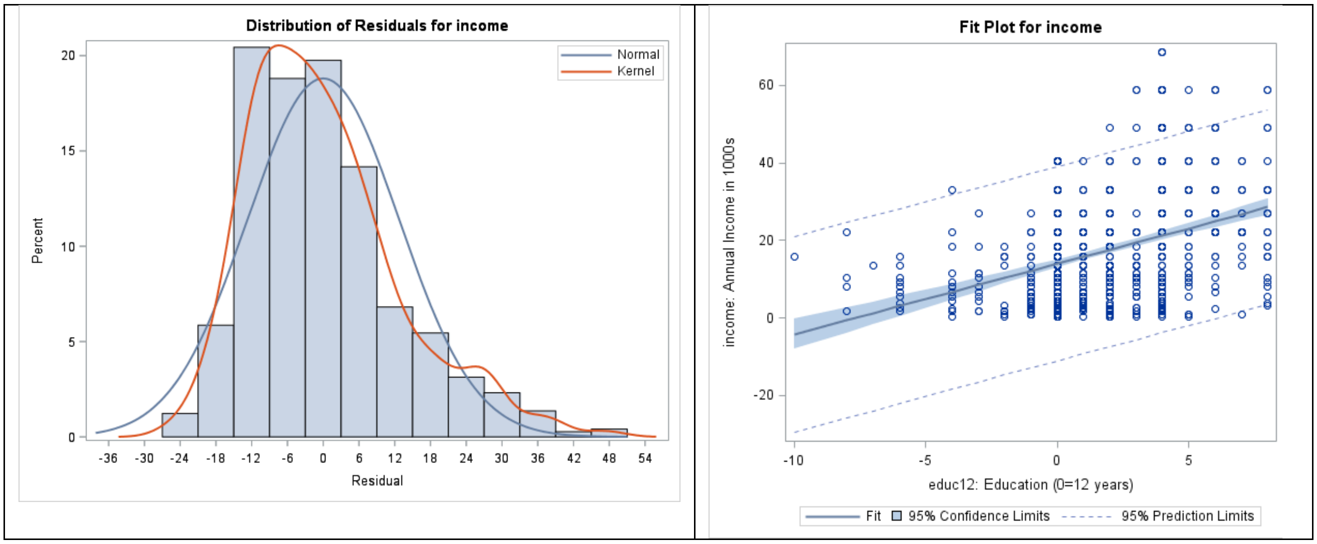
```
ESTIMATE "Pred Income  8 years (educ12=-4)" intercept 1 educ12 -4;
ESTIMATE "Pred Income 12 years (educ12= 0)" intercept 1 educ12  0;
ESTIMATE "Pred Income 16 years (educ12= 4)" intercept 1 educ12  4;
ESTIMATE "Pred Income 20 years (educ12= 8)" intercept 1 educ12  8;
```

This is the extra table of linear combinations of the fixed effects created by SAS ESTIMATEs:

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

Last I checked, SPSS GLM does not have these SAS ESTIMATE commands that provide linear combinations of model parameters, which is one of the reasons I don't teach using SPSS. However, you can also estimate GLMs using SPSS MIXED, in which the /TEST subcommand works exactly liked ESTIMATE in SAS.

Part of plots from SAS PROC GLM—residuals are not yet normal with constant variance:



Using Centered Education Predictor 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
// In 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 at  8 years (educ12=-4)
lincom _cons*1 + educ12*0    // Pred Income at 12 years (educ12= 0)
lincom _cons*1 + educ12*4    // Pred Income at 16 years (educ12= 4)
lincom _cons*1 + educ12*8    // Pred Income at 18 years (educ12= 8)
```

```
      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
```

This is the regular table of fixed effects estimated directly by the model:

```
------------------------------------------------------------------------------
      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
------------------------------------------------------------------------------
```

Interpret β_0 = intercept:

Interpret β_1 = slope of education−12:

The LINCOM commands provide an example of how to compute predicted values for the outcome given any value(s) of the predictor(s). Model: $Income_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$

```
. lincom _cons*1 + educ12*-4   // Pred Income at  8 years (educ12=-4)
( 1)  - 4*educ12 + _cons = 0
------------------------------------------------------------------------------
      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 at 12 years (educ12= 0)
 ( 1)  _cons = 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 at 16 years (educ12= 4)
 ( 1)  4*educ12 + _cons = 0
------------------------------------------------------------------------------
      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 at 18 years (educ12= 8)
 ( 1)  8*educ12 + _cons = 0
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   28.58823   1.105747    25.85   0.000     26.41742    30.75905
------------------------------------------------------------------------------
```

```
// To make regression plots given in SAS
display as result "STATA Regression Line with CI for mean (stdp option)"
graph twoway lfitci income educ12, stdp || scatter income educ12

display as result "STATA Regression Line with CI for individual (stdf option)"
graph twoway lfitci income educ12, stdf || scatter income educ12
```

Standardized Solution using Centered Education Predictor in SAS:

```
TITLE1 "SAS GLM Predicting Income from Centered Education";
TITLE2 "Using REG instead of GLM to get standardized Effects";
PROC REG DATA=work.Example3;
     MODEL income = educ12 / STB; * STB option gives standardized solution;
RUN; QUIT; TITLE1; TITLE2;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized 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**

Standardized Solution using Centered Education Predictor in STATA:

In the standardized solution, fixed slopes are given in a correlation metric (−1 to 1).

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

Last model: Income predicted by binary marital status (1=unmarried, 2=married)…

Given that no one in this sample had marital status = 0, let's create a new centered version by subtracting 1 so that the groups = 0 or 1, which is known as "dummy coding":
$\boldsymbol{Income_i = \beta_0 + \beta_1(Marry01_i) + e_i}$

Using Centered (Dummy-Coded) Marry01 Predictor in SAS:

```
* Center marry predictor so that 0 is meaningful;
DATA work.Example3; SET work.Example3;
     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.Example3 NAMELEN=100;
     MODEL income = marry01 / SOLUTION ALPHA=.05 CLPARM;
* ESTIMATEs below request predicted outcome means for each group;
     ESTIMATE "Income for Unmarried (marry01=0)" intercept 1 marry01 0; * Beta0;
     ESTIMATE "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.

This is the regular table of fixed effects estimated directly by the model :

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**

Predicted income unmarried (marry01=0): $\hat{y}_i = \mathbf{14.45} + \mathbf{6.22}(\mathbf{0}) = \mathbf{14.45}$
Predicted income unmarried (marry01=1): $\hat{y}_i = \mathbf{14.45} + \mathbf{6.22}(\mathbf{1}) = \mathbf{20.67}$

```
* ESTIMATEs below request predicted outcome means for each group;
     ESTIMATE "Income for Unmarried (marry01=0)" intercept 1 marry01 0; * Beta0;
     ESTIMATE "Income for Married   (marry01=1)" intercept 1 marry01 1; * Beta0+Beta1;
```

This is the extra table of linear combinations of the fixed effects created by SAS ESTIMATEs:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Income for Unmarried (marry01=0)	14.4454345	0.67488958	21.40	<.0001	13.1204845	15.7703845
Income for Married (marry01=1)	20.6690579	0.73250910	28.22	<.0001	19.2309886	22.1071271

Interpret $\boldsymbol{\beta_0}$ = intercept:

Interpret $\boldsymbol{\beta_1}$ = slope of marry01:

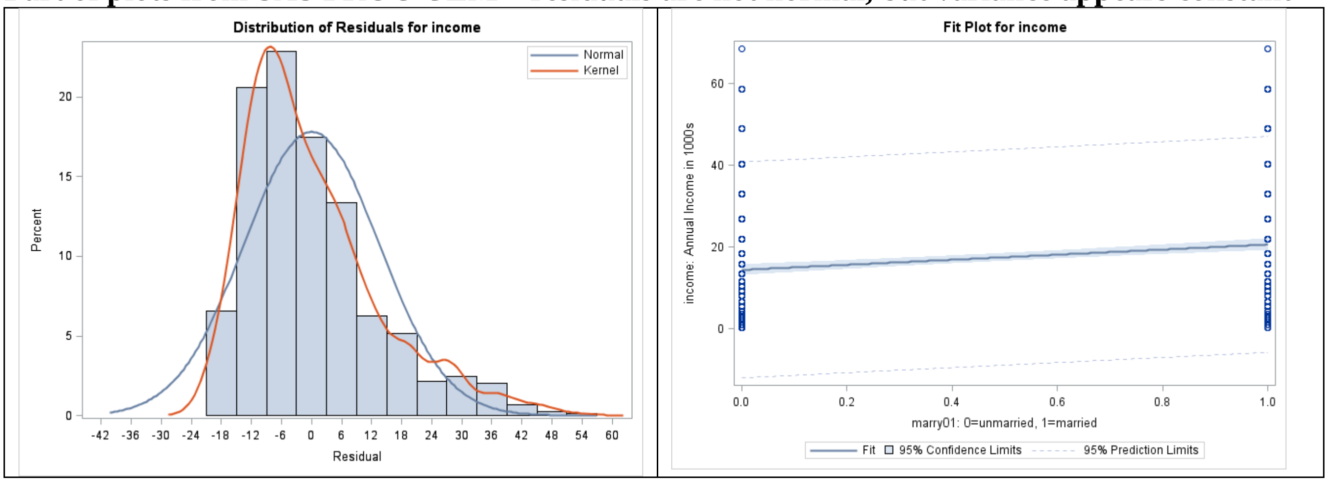
To get a Cohen's d effect size for the mean income difference between unmarried and married persons, we can calculate d from the t test-statistic: $d = \frac{2t}{\sqrt{DF_{den}}} = \frac{2*6.25}{\sqrt{732}} = 0.46$ → the mean income difference is about 0.46 standard deviations higher for married than unmarried persons.

```
* Compute d effect size for marry01 from t test-statistic;
DATA work.MakeD;
     d=2*6.25/SQRT(732);
RUN;
* Print results of d computation;
PROC PRINT NOOBS DATA=work.MakeD;
RUN;
```

The code on the left makes a new dataset, creates a new variable d that holds the result of the formula, and then PROC PRINT prints that new dataset to the output.

```
     d
0.46201
```

Part of plots from SAS PROC GLM—residuals are not normal, but variance appears constant



Using Centered (Dummy-Coded) Marry01 Predictor in STATA:

```
// Center 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 // Income for Unmarried (Marry01=0) = Beta0
lincom _cons*1 + marry01*1 // Income for Married   (Marry01=1) = Beta0 + Beta1
```

```
      Source |       SS           df       MS      Number of obs   =       734
-------------+----------------------------------   F(1, 732)       =     39.04
       Model |  7060.10161         1  7060.10161   Prob > F        =    0.0000
    Residual |   132363.13       732  180.823948   R-squared       =    0.0506
-------------+----------------------------------   Adj R-squared   =    0.0493
       Total |  139423.232       733  190.209048   Root MSE        =    13.447
```

This is the regular table of fixed effects estimated directly by the model:

```
------------------------------------------------------------------------------
      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
------------------------------------------------------------------------------
```

The LINCOM commands provide an example of how to compute predicted values for the outcome given any value(s) of the predictor(s). Model: $Income_i = \beta_0 + \beta_1(Marry01) + e_i$

```
. lincom _cons*1 + marry01*0 // Income for Unmarried (Marry01=0) = Beta0
 ( 1)  _cons = 0
------------------------------------------------------------------------------
      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 // Income for Married   (Marry01=1) = Beta0 + Beta1
 ( 1)  marry01 + _cons = 0
------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   20.66906   .7325091    28.22   0.000     19.23099    22.10713
------------------------------------------------------------------------------
```

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: $d = \frac{2t}{\sqrt{DF_{den}}} = \frac{2*6.25}{\sqrt{732}} = 0.46$ → the mean income difference is about 0.46 standard deviations higher for married than unmarried persons.

```
// Compute d effect size for marry01 from t test-statistic
display 2*6.25/sqrt(732)

. display 2*6.25/sqrt(732)
.46201329
```

<u>Example Results Section:</u>
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 4].

We then examine 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* = 0.46.

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* = 20.67k, SE = 0.73) was significantly higher than for unmarried persons (*M* = 14.45k, SE = 0.67).