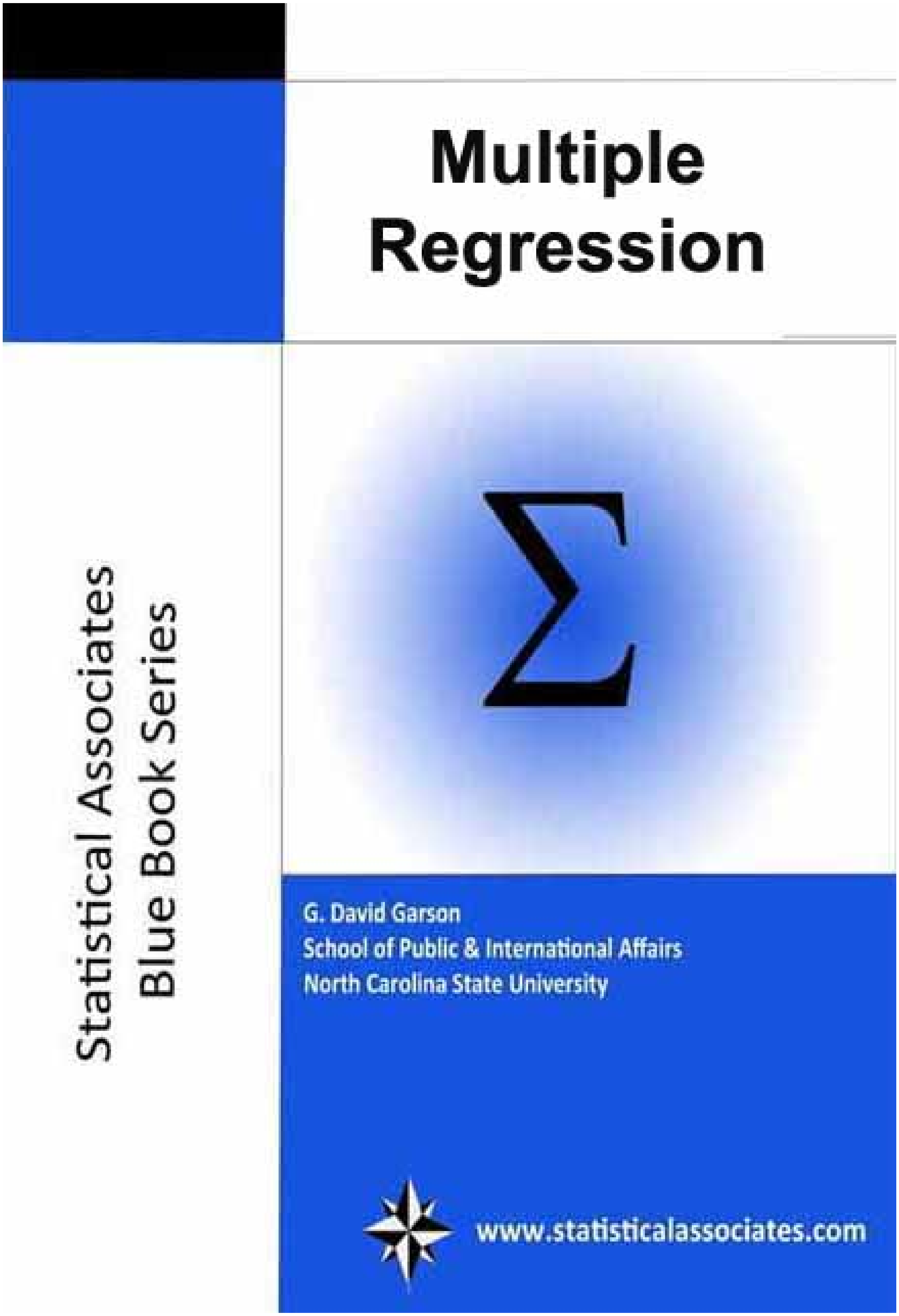


Multiple Regression

Statistical Associates
Blue Book Series

G. David Garson
School of Public & International Affairs
North Carolina State University

www.statisticalassociates.com

@c 2014 by G. David Garson and Statistical Associates Publishing. All rights reserved worldwide in all media. No permission is granted to any user to copy or post this work in any format or any media.

ISBN-10: 1626380317

ISBN-13: 978-1-62638-031-8

The author and publisher of this eBook and accompanying materials make no representation or warranties with respect to the accuracy, applicability, fitness, or completeness of the contents of this eBook or accompanying materials. The author and publisher disclaim any warranties (express or implied), merchantability, or fitness for any particular purpose. The author and publisher shall in no event be held liable to any party for any direct, indirect, punitive, special, incidental or other consequential damages arising directly or indirectly from any use of this material, which is provided "as is", and without warranties. Further, the author and publisher do not warrant the performance, effectiveness or applicability of any sites listed or linked to in this eBook or accompanying materials. All links are for information purposes only and are not warranted for content, accuracy or any other implied or explicit purpose. This eBook and accompanying materials is © copyrighted by G. David Garson and Statistical Associates Publishing. No part of this may be copied, or changed in any format, sold, or used in any way under any circumstances other than reading by the downloading individual.

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: gdavidgarson@gmail.com
Web: www.statisticalassociates.com

Table of Contents

Page numbering works in the full edition. This is a preview edition of the first 30 pages.

Multiple Regression

Overview

Multiple regression, a time-honored statistical technique going back to Karl Pearson's use of it in 1908, is employed to account for the variance in a continuous dependent variable, based on linear combinations of interval, dichotomous, or dummy independent variables. Often called OLS regression because of its reliance on ordinary least squares estimation, multiple regression can establish whether a set of independent variables explains a proportion of the variance in a dependent variable (through $R^{2,}$ which is percent of variance explained) at a significant level (through a significance test of R^2), and can establish the relative predictive importance of the independent variables (by comparing beta weights, which are standardized regression coefficients).

In multiple regression…

- Power terms can be added as independent variables to explore curvilinear effects.
- Cross-product terms can be added as independent variables to explore interaction effects.
- The researcher can test the significance of the difference between two models, with and without a given predictor variable, to determine if adding an independent variable to the model would improve the model significantly.
- Using hierarchical regression (a.k.a. “block regression”, not to be confused with hierarchical linear regression, which is multilevel regression), the researcher can see how much variance in the dependent variable can be explained by one or a set of new independent variables, over and above that explained by an earlier set.
- Finally, the parameter estimates (the b coefficients and the constant) can be used to construct a prediction equation and generate predicted scores of the dependent variable for further analysis.

A multiple regression equation takes the form $y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$. The b parameter estimates are the regression coefficients, representing the amount the dependent variable (y) changes when the corresponding independent variable changes 1 unit, controlling for other variables in the model. The c is the constant, indicating where the regression line intercepts the y axis, representing the average magnitude of the dependent variable when all the independent variables are held to 0. The the data have been centered, this will be controlling all other variables at their means. The standardized version of the b coefficients are the beta weights, and the ratio of the beta weights is often interpreted as the ratio of the relative predictive power of the independent variables. Associated with multiple regression is R^2 (a.k.a multiple correlation), which is the percent of variance in the dependent variable explained collectively by all of the independent variables in the model.

Multiple regression shares all the assumptions of correlation: linearity of relationships, the same level of relationship throughout the range of the dependent variable (homoscedasticity), interval or near-interval measurement level, absence of outliers, and data whose range is not truncated. In addition, it is important that the model being tested is correctly specified. The exclusion of important causal variables or the inclusion of extraneous variables can change markedly the b coefficients and beta weights and hence the interpretation of the importance of the independent variables.

There are many of alternatives to ordinary least squares (OLS) regression. These are treated in separate volumes of the Statistical Associates "Blue Book" series. Some of these topics are listed alphabetically below.

- Cox regression may be used to analyze time-to-event as well as proximity, and preference data. See the separate Statistical Associates “Blue Book” volume on “Cox Regression”, listed at the end of this volume.
- Curve estimation lets the researcher explore how linear regression compares to nonlinear models, useful for exploring which statistical procedures and models may be appropriate for relationships in one's data. See the separate Statistical Associates “Blue Book” volume on “Curve Estimation & Nonlinear Regression”.
- Discriminant function analysis. is used when the dependent variable is a dichotomy but other assumptions of multiple regression can be met, making it more powerful than the logistic alternatives. See the separate Statistical Associates “Blue Book” volume on “Discriminant Function Analysis”.

- The general linear modeling (univariate GLM) is primarily used for analysis of variance designs but also may be used to implement multiple regression. See the separate Statistical Associates "Blue Book" volume on "GLM Univariate, ANOVA, and ANCOVA".
- The general linear model (multivariate GLM). Multivariate GLM can implement regression models with multiple dependent variables. See the separate Statistical Associates "Blue Book" volume on "GLM Multivariate, MANOVA, and MANCOVA".
- Generalized linear models (GZLM) is the generalization of linear modeling to a form covering almost any dependent distribution with almost any link function, thus supporting linear regression, Poisson regression, gamma regression, and many others. See the separate Statistical Associates "Blue Book" volume on "Generalized Linear Models: Generalized Estimating Equations".
- Generalized estimating equations (GEE) is similar to GZLM but hands repeated measures and other non-independent data. See above.
- Linear mixed models (LMM), also called hierarchical linear models, implement regression in the context of multilevel data and linear effects of higher levels on lower levels. See Garson (2013).
- Generalized linear mixed models (GLMM) implement regression for multilevel data, supporting a variety of nonlinear link functions.
- Logistic regression is used for dichotomous and multinomial dependent variables, implemented with stand-alone logistic procedures or through generalized linear models. See the separate Statistical Associates "Blue Book" volume on "Logistic Regression: Binomial & Multinomial".
- Logit regression is an equivalent to logistic regression, using log-linear techniques to predict one or more categorical dependent variables. See the separate Statistical Associates "Blue Book" volume on "Log-linear Models".
- Multinomial regression handles research where the dependent variable is categorical. See the separate Statistical Associates "Blue Book" volume on "Logistic Regression: Binomial & Multinomial".
- Nonlinear regression is used when the model is inherently nonlinear (nonlinearities cannot be dealt with using link functions in generalized linear models or by power and other transformations in general linear models, including regression). See the separate Statistical Associates "Blue Book" volume on "Curve Estimation & Nonlinear Regression".

- Ordinal regression has more power than multinomial regression when the levels of a categorical dependent variable are ordered. See the separate Statistical Associates "Blue Book" volume on "Ordinal Regression".
- Partial least squares regression, which merges regression and factor analysis techniques, may be used even with small datasets to predict a set of response variables from a set of independent variables. See the separate Statistical Associates "Blue Book" volume on "Partial Least Squares Regression and Structural Equation Models".
- Partial proportional odds (PPO) regression is an alternative to ordinal regression when the parallel lines test fails. See the separate Statistical Associates "Blue Book" volume on "Ordinal Regression".
- Poisson regression is used for count data in survival (event history) analysis and other techniques, implemented with stand-alone general loglinear procedures and through generalized linear models. See above.
- Two-stage least squares regression (2SLS) may be used when one or more independent variables are correlated with the error term, when omitted variables may be treated using instrumental variables, and when the model involves interdependence among predictor variables and thus is recursive. See the separate Statistical Associates "Blue Book" volume on "Two-Stage Least Squares Regression".
- Weighted least squares (WLS) regression may be used when the OLS regression assumption of homoscedasticity has been violated. See the separate Statistical Associates "Blue Book" volume on "Weighted Least Squares Regression".

Data examples in this volume

The example datasets used in this volume are listed below in order of use, with versions for SPSS (.sav), SAS (.sas7bdat), and Stata (.dta).

Except where otherwise noted, examples in this volume are from a modified version of the GSS93subset.sav General Social Survey data file originally supplied in the SPSS Samples directory.

- Click here to download GSS93subset.sav for SPSS.
- Click here to download GSS93subset.sas7bdat for SAS.
- Click here to download GSS93subset.dta for Stata.

A section below on interaction effects uses the "cars" dataset in a version modified from that found in the SPSS Samples directory. Click here to download cars.sav for SPSS.

- Click here to download cars.sav for SPSS.
- Click here to download cars.sas7bdat for SAS.
- Click here to download cars.dta for Stata.

A section below on quantile regression uses the "auto" dataset in a version based on that provided by Stata.

- Click here to download auto.sav for SPSS.
- Click here to download auto.sas7bdat for SAS.
- Click here to download auto.dta for Stata.

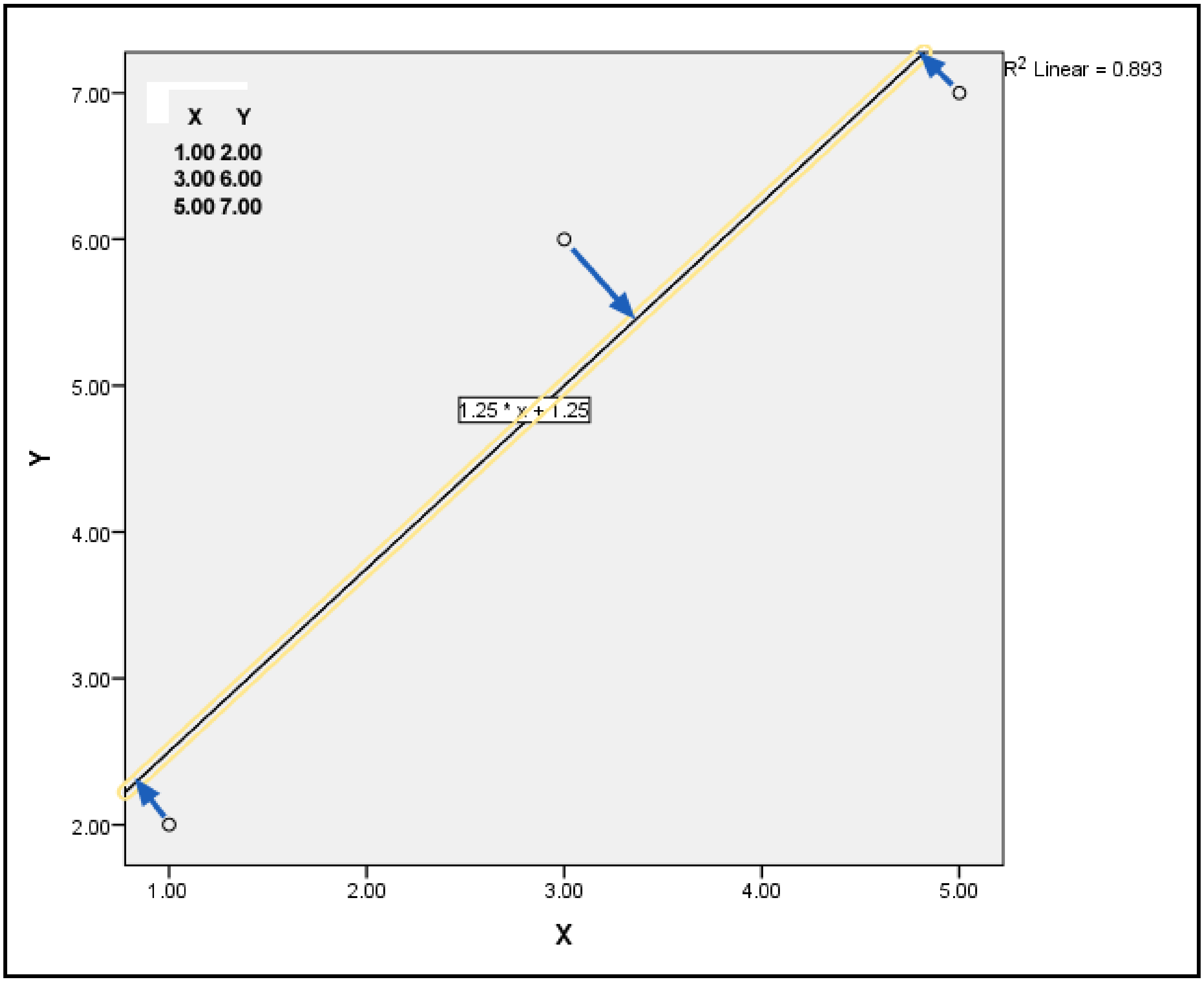
A section below on difference-in-difference regression uses the "kielmcclainsubset" dataset in a version based on that provided by Stata.

- Click here to download kielmcclainsubset.sav for SPSS.
- Click here to download kielmcclainsubset.sas7bdat for SAS.
- Click here to download kielmcclainsubset.dta for Stata.

Key Terms and Concepts

OLS estimation

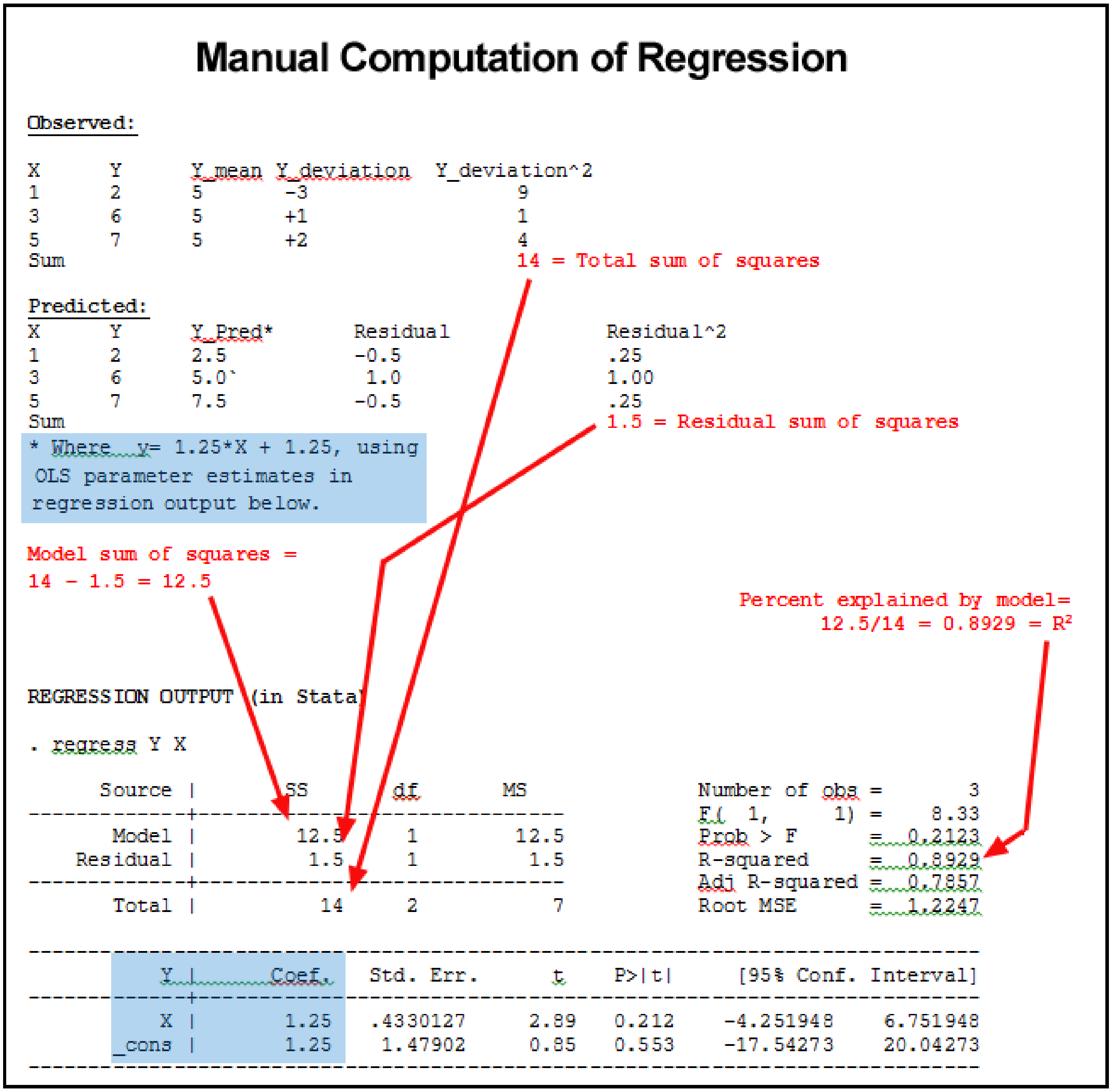
OLS stands for ordinary least squares. The name is derived from the criterion used to draw the best fit regression line: a line such that the sum of the squared deviations of the distances of all observed points to the line is minimized. In the illustration below, Y is regressed on X for three data points. The OLS algorithm creates a prediction equation (here, Y = 1.25*X + 1.25) such that the squared distances to the prediction line are minimized.



The regression equation

The OLS regression equation takes the form $Y = b_1*x_1 + b_2*x_2 + c + e$, where Y is the raw dependent, the b's are the regression coefficients for the corresponding x (independent) terms, where c is the constant or intercept, and e is the error term reflected in the residuals. Sometimes this is expressed more simply as $y = b_1*x_1 + b_2*x_2 + c$, where y is the estimated dependent variable and c is the constant (which includes the error term). Equations such this, with no interaction effects (see below), are called *main effects models*.

In the figure below, key regression statistics are computed based on the three data points in the figure above and on the OLS-derived equation, $Y = 1.25*X + 1.25$.



Manual Computation of Regression

```
Observed:

X       Y       Y_mean  Y_deviation   Y_deviation^2
1       2       5        -3                   9
3       6       5        +1                   1
5       7       5        +2                   4
Sum                                           14 = Total sum of squares

Predicted:
X       Y       Y_Pred*        Residual            Residual^2
1       2       2.5            -0.5                .25
3       6       5.0`            1.0                1.00
5       7       7.5            -0.5                .25
Sum                                                1.5 = Residual sum of squares
* Where  y= 1.25*X + 1.25, using
 OLS parameter estimates in
 regression output below.

Model sum of squares =
14 - 1.5 = 12.5
                                                    Percent explained by model=
                                                        12.5/14 = 0.8929 = R²

REGRESSION OUTPUT (in Stata)

. regress Y X

      Source |       SS       df       MS              Number of obs =       3
-------------+------------------------------           F(  1,     1) =    8.33
       Model |        12.5     1        12.5           Prob > F      =  0.2123
    Residual |         1.5     1         1.5           R-squared     =  0.8929
-------------+------------------------------           Adj R-squared =  0.7857
       Total |          14     2           7           Root MSE      =  1.2247

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |       1.25   .4330127     2.89   0.212    -4.251948    6.751948
       _cons |       1.25    1.47902     0.85   0.553    -17.54273    20.04273
------------------------------------------------------------------------------
```

The simple example above illustrates some of the key components of regression:

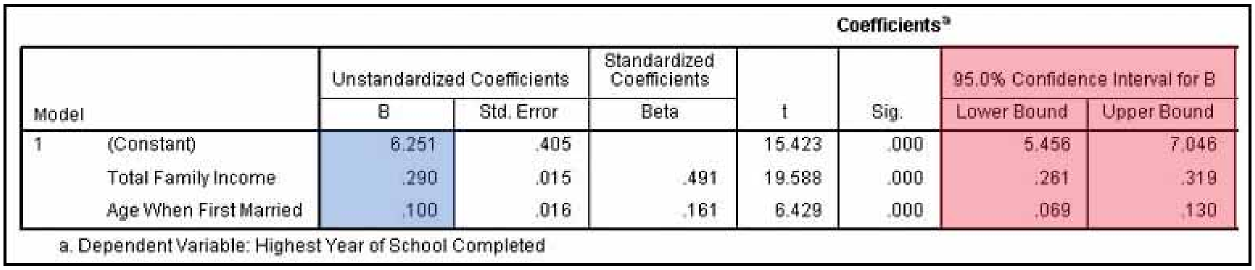
1. *OLS parameter estimates*: Shown in blue in the figure above, the b coefficient for X is 1.25 and by coincidence the constant is also 1.25. The regression (prediction) equation is therefore Y = 1.25*X + 1.25.
2. *Total sum of squares:* This is the sum of the squared deviations of the Y (dependent) variable from its mean.
3. *Residual sum of squares*: This is the sum of the squared residuals, where residuals are the differences between the observed Y and the predicted Y.

4. *Model sum of squares:* The total sum of squares minus the residual sum of squares is the model sum of squares.
5. *R^2:* R-square is the model sum of squares as a percentage of the total sum of squares and is interpreted as the percent of variance in Y explained by the regression model. Here, 89.29% of the variance in Y is explained by X.

Second example. In the figure below, using SPSS output, the highest year of school completed (educ) is predicted from total family income (income91) and age when first married (agewed). Using the parameter estimates shown in blue, the regression equation is:

educ = .29*income91 + .10*agewed + 6.251.

If 0 is not within the 95% confidence limits (shown in red below), then the regression coefficient is significant at at least the .05 significance level, as both are for this example.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.251	.405		15.423	.000	5.456	7.046
	Total Family Income	.290	.015	.491	19.588	.000	.261	.319
	Age When First Married	.100	.016	.161	6.429	.000	.069	.130

a. Dependent Variable: Highest Year of School Completed

Dependent variable

The dependent variable is the predicted variable in the regression equation (educ in the example above). Also called response, outcome, or criterion variables, dependent variables are assumed to be continuous, interval variables, though it is common to see binary or ordinal independent variables in linear regression. Use of binary variables as dependent variables is no longer acceptable since such variables cannot meet regression's normal distribution assumption - logistic regression is commonly used instead. Likewise, use of ordinal variables as dependent variables is now derogated in favor of ordinal regression and proportional odds models (treated in the “Ordinal Regression” title of the Statistical Associates “Blue Book” series.

Independent variables

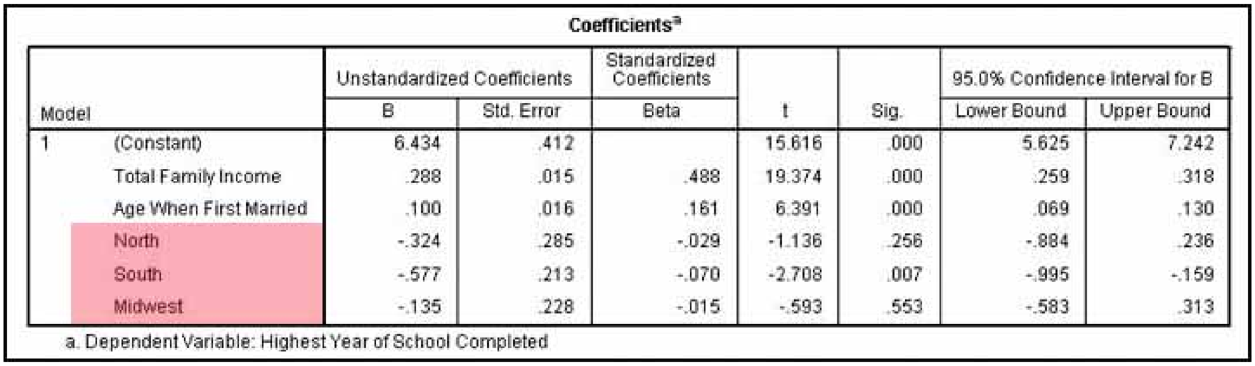
Independent variables are the predictor variables in the regression equation (income91 and agewed in the example above). Predictor variables usually are continuous variables. It is, however, common to see the use of ordinal predictor variables in linear regression even though this violates the assumptions of OLS regression. Use of ordinal predictor variables with fewer than five levels is particularly derogated. It is acceptable to use binary predictor variables and to transform nominal and ordinal categorical variables into sets of dummy variables (coded 0, 1, and leaving one level out as the reference category to avoid perfect multicollinearity). See the discussion in the "Assumptions" section below.

Dummy variables

Dummy variables are a way of adding the values of a nominal or ordinal variable to a regression equation. The standard approach to modeling categorical variables is to include the categorical variables in the regression equation by converting each level of each categorical variable into a variable of its own, usually coded 0 or 1. For instance, the categorical variable "region" may be converted into dummy variables such as "East," "West," "North," or "South." Typically "1" means the attribute of interest is present (ex., South = 1 means the case is from the region South).

Once a set of dummy variables is created, if we know an observation's value on all the levels of a categorical variable except one, that last one is determined. We have to leave one of the levels out of the regression model to avoid perfect multicollinearity (a.k.a. singularity or redundancy), which will prevent a solution. For example, we may leave out "West" to avoid singularity. The reference level should be the level of greatest interest or at least a level with known characteristics. Letting a residual category (e.g., “Other”) be the reference category undermines the meaningfulness of reference statements.

The omitted category is the reference category because b coefficients must be interpreted with reference to it. In the example output below, three regions (North, South, Midwest, where West is the left-out reference category) are added as dummy predictors. Since West has the highest mean score on the dependent (highest year of school completed), all the b coefficients are negative, meaning that if a subject is in any other region than West, a lower value is predicted. Only being in the South is significantly lower than being from the West (the reference category), however. Being in the South predicts 0.577 fewer years of education than being in the West, controlling for other variables in the model. See further discussion below.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.434	.412		15.616	.000	5.625	7.242
	Total Family Income	.288	.015	.488	19.374	.000	.259	.318
	Age When First Married	.100	.016	.161	6.391	.000	.069	.130
	North	-.324	.285	-.029	-1.136	.256	-.884	.236
	South	-.577	.213	-.070	-2.708	.007	-.995	-.159
	Midwest	-.135	.228	-.015	-.593	.553	-.583	.313

a. Dependent Variable: Highest Year of School Completed

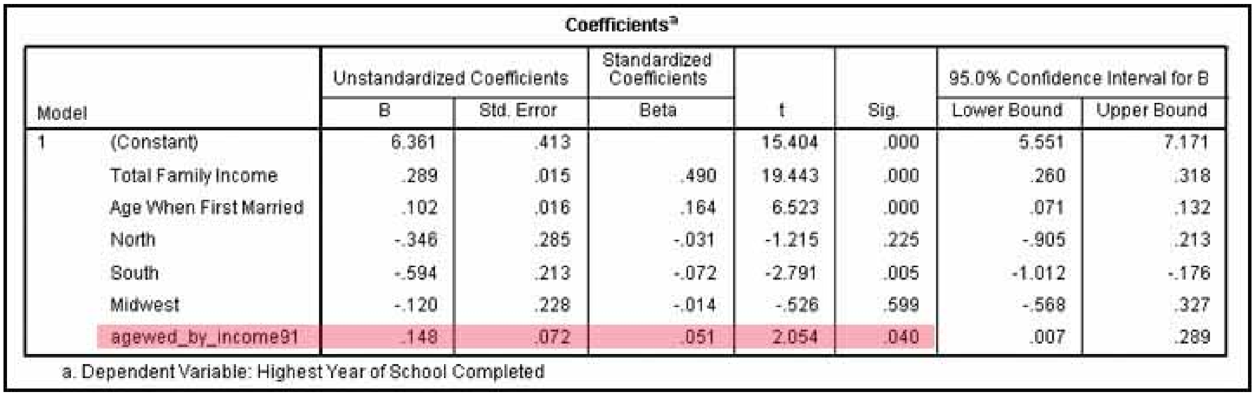
Note that dummy variables are not used as dependent variables in a normal regression model. See further discussion in the FAQ section below, including alternative methods of coding categorical dummy variables.

Interaction effects

Interactions

Interaction effects are sometimes called *moderator effects* because the interacting third variable which changes the relation between two original variables moderates the original relationship. However, an interaction effect is a special type of moderator effect which has an impact going beyond simple main effects of individual variables. For instance, the impact of being African American or of being a woman may be significant, but there may be an additional impact of being an African American woman over and above the main effects of race and gender.

Interaction terms may be added to the model to incorporate the joint effect of two variables (ex., race and gender) on a dependent variable (ex., income) over and above their separate effects. The researcher adds interaction terms to the model as crossproducts of the standardized independent variables and/or dummy independent variables, typically placing them after the simple "main effects" independent variables. In the example below, the interaction term "agewed_by_income91" is added to the previously discussed regression model. It is found to be significant at the .04 level, meaning that there is a combined effect of age when first wed with family income over and beyond that of either agewed or income individually.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.361	.413		15.404	.000	5.551	7.171
	Total Family Income	.289	.015	.490	19.443	.000	.260	.318
	Age When First Married	.102	.016	.164	6.523	.000	.071	.132
	North	-.346	.285	-.031	-1.215	.225	-.905	.213
	South	-.594	.213	-.072	-2.791	.005	-1.012	-.176
	Midwest	-.120	.228	-.014	-.526	.599	-.568	.327
	agewed_by_income91	.148	.072	.051	2.054	.040	.007	.289

a. Dependent Variable: Highest Year of School Completed

Margin plots, discussed below in the FAQ section, can reveal interaction effects graphically, where they appear in the form on non-parallel lines which may cross.

For further reading, a leading reference for interaction effects in regression is Aiken & West (1991).

Centering

Crossproduct interaction terms may be highly correlated (multicollinear) with the corresponding simple independent variables in the regression equation, creating problems with assessing the relative importance of main effects and interaction effects. Centering the predictor variables (or standardizing, which incorporates centering) tends to reduce multicollinearity and is therefore recommended. Note also that there are alternatives to the crossproduct approach to analyzing interactions (see discussion below). See also general discussion of centering below.

Significance of interaction effects

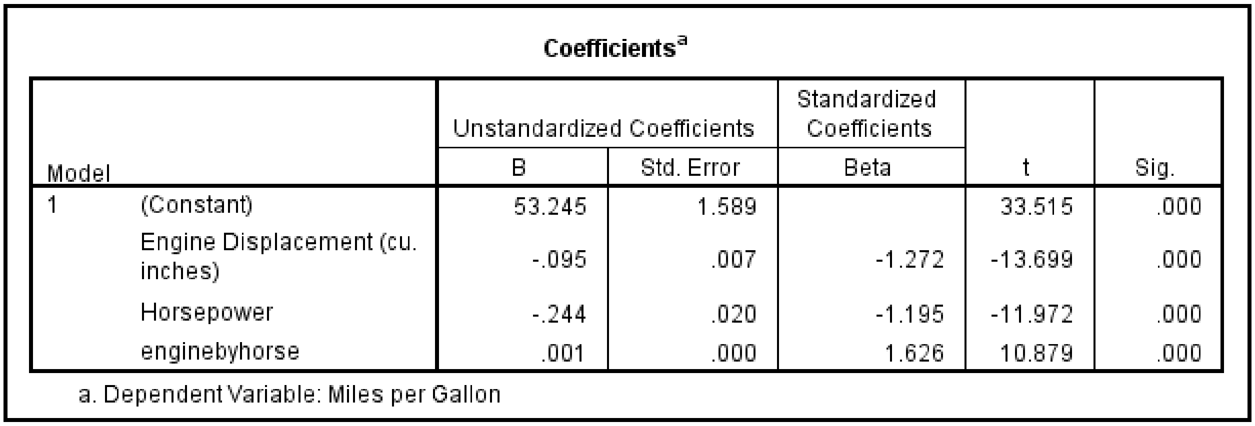
The significance of an interaction effect is the same as for any other variable, except in the case of a set of dummy variables representing a single categorical variable. When an ordinal variable has been entered as a set of dummy variables, the interaction of another variable with the ordinal variable will involve multiple interaction terms. In this case the F-test of the significance of the interaction of the two variables is the significance of the change of R-square of the equation with the interaction terms and the equation without the set of terms associated with the categorical variable.

Interaction terms with categorical dummies

To create an interaction term between a categorical variable and a continuous variable, first the categorical variable is dummy-coded, creating (k - 1) new variables, one for each level of the categorical variable except the omitted reference category. The continuous variable is multiplied by each of the (k - 1) dummy variables. The terms entered into the regression include the continuous variable, the (k - 1) dummy variables, and the (k - 1) cross-product interaction terms. Also, a regression is run without the interaction terms. The R-squared difference measures the effect of the interaction. The beta weights for the interaction terms in the regression which includes the interaction terms measure the relative predictive power of the effects of the continuous variable given specific levels of the categorical variable. It is also possible to include both interaction terms and power terms in the model by multiplying the dummies by the square of the continuous variable, but this can lead to an excessive number of terms in the model. One approach to dealing with this is to use stepwise regression's F test as a criterion to stop adding terms to the model.

Plotting interactions through simple slope analysis

Simple slope analysis serves as an exploration tool for understanding interaction effects more deeply. This subsection uses the “cars” dataset (see above), containing attributes of U.S. and foreign cars. Consider the table below, in which miles per gallon (MPG) is predicted from engine displacement in cubic inches (engine), from horsepower, and the interaction of the two (enginebyhorse). Both main effects and the interaction effect are all significant at better than the .001 level.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	53.245	1.589		33.515	.000
	Engine Displacement (cu. inches)	-.095	.007	-1.272	-13.699	.000
	Horsepower	-.244	.020	-1.195	-11.972	.000
	enginebyhorse	.001	.000	1.626	10.879	.000

a. Dependent Variable: Miles per Gallon

We use the following notation:

DV	miles per gallon (MPG), the dependent variable
X_1	engine
B_1	regression coefficient for engine = -.095
X_2	horsepower
B_2	regression coefficient for horsepower = -.244
X_1X_2	enginebyhorse, which is the engine*horsepower interaction
B_3	the regression coefficient for the interaction effect = .001
C	the constant = 53.245
Y'	the estimated value of MPG

The regression equation is:

(1) $Y' = C + B_1X_1 + B_2X_2 + B_3(X_1X_2)$

When there is an interaction effect, such as one affecting the relationship of engine size (X1) and MPG (the DV), then the engine size slope (b1) will be different for different values of the interacting variable, here horsepower (X2). We can test this by seeing whether B1 at high levels of X2 differs from B1 at low levels of X2.

While any pair of values of X2 might be spotlighted (hence simple slope analysis may be called a form of “spotlight analysis), Cohen et al. (2003) state that plus or minus on standard deviation of the mean may be used as high and low values, absent any other theoretical basis for selecting particular high and low values. For X2 = horsepower, the mean is 104.93; the standard deviation is 38.522; plus 1 sd is 143.452; and -1 sd is 66.408.

To better obtain the regression formula for Y' at high and low values of X_2, the regression formula above may be transformed by simple algebra as follows:

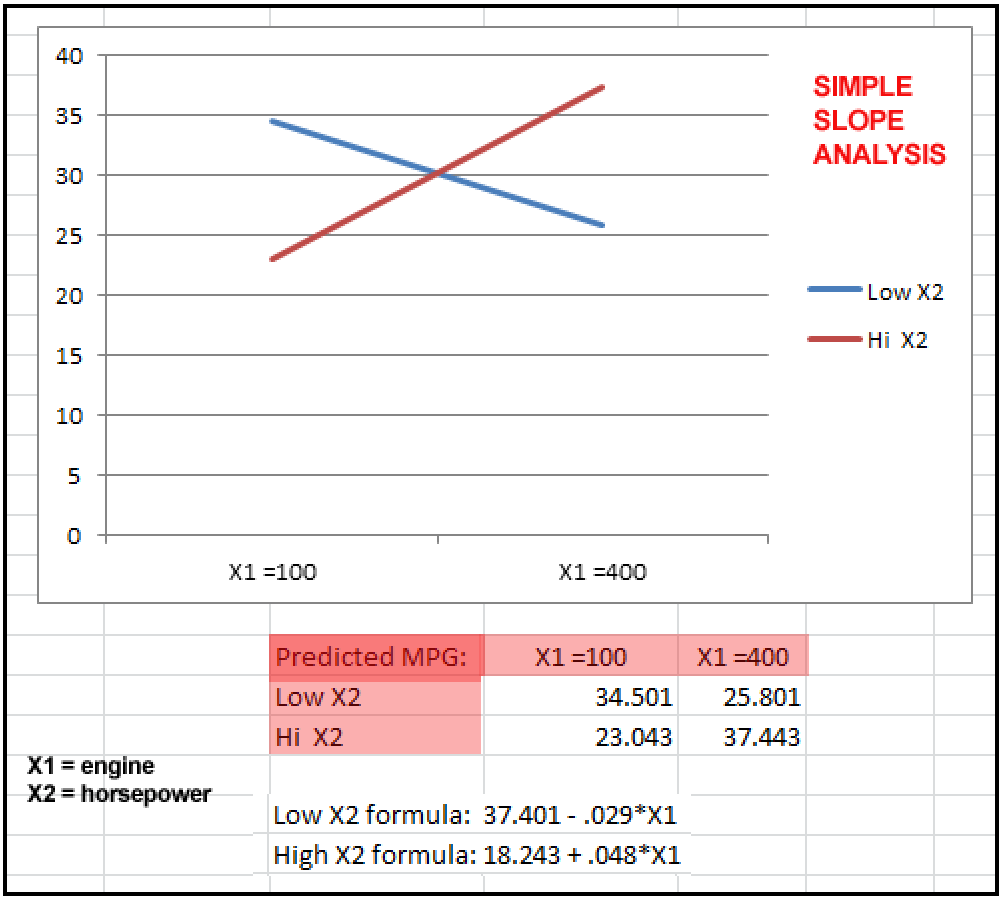
(2) $Y' = C + B_2X_2 + B_1X_1 + B_3(X_1X_2)$

(3) $Y' = (C + B_2X_2) + (B_1 + B_3X_2)*X_1$

Using formula (3) above, we may compute the regression formulas for Y' at high and low values of X_2.

High: $Y' = (53.245+(-.244)(143.452)) + (-.095 + .001(143.452))*X_1$
$= (53.245 – 35.002) + (-.095 +.143) *X_1$
(4) $= 18.243 + .048*X_1$
Low: $Y' = (53.245+(-.244)(66.408)) + (-.095 + .001(66.408))*X_1$
$= (53.245 – 16.204) – (-.095 +.066) *X_1$
(5) $= 37.401 -.029*X_1$

We immediately see that when X_2 (horsepower) is high, the slope of X_1 (engine displacement) is +.048, but when X_2 is low, the slope of X_1 is -.029. That the slope of engine displacement is different at different levels of horsepower shows an interaction effect involving horsepower. This is shown graphically if equations (4) and (5) are plotted. S below (using Microsoft Excel), where the crossed slopes indicate an interaction effect. This is confirmed in the regression table above, where the enginebyhorse interaction is significant.



Predicted MPG:	X1 =100	X1 =400
Low X2	34.501	25.801
Hi X2	23.043	37.443

X1 = engine
X2 = horsepower

Low X2 formula: 37.401 - .029*X1
High X2 formula: 18.243 + .048*X1

For more on simple slope analysis, see Aiken & West (1991) and Tabachnick & Fidell (2013:158-161).

Separate regressions

An alternative but little-used approach to interactions is to run separate regressions for each level of the interacting variable. There are, however, objections to this approach. While separate regressions are feasible for a single variable such as gender, running a male and a female regression, it is not the best approach for two reasons. In practical terms, if there are multiple categorical variables each with multiple categories (levels), the number of needed regressions may become unwieldy and the sample size for individual regressions may become too small. Smaller sample size means the researcher will lose statistical power compared to one overall regression. Loss of statistical power means the researcher will be more likely to make Type II errors (false negatives, thinking there is no relationship when in reality there is).

Predicted values

Predicted values, also called fitted values, are the values of each case based on using the regression equation for all cases selected for analysis. SPSS, SAS, and Stata all support generating fitted values and saving these as new variables to the researcher's working dataset, which can then be saved to file with the command File > Save As.

SPSS

Saving estimates is accomplished in SPSS under the "Output" button. SPSS will save the predicted values under the term PRED to refer to predicted values and ZPRED to refer to standardized predicted values. Click the Save button in SPSS to add and save these as new variables in your dataset.

SAS

In SAS, saving is accomplished by defining a library and using the OUTPUT statement, as in the syntax below:

```
LIBNAME in "C:\Data";
```

PROC REG syntax supports an OUTPUT statement.The syntax below creates a file called mydata.sas7bdat in the C:\Data folder. This file contains all the variables from the original dataset plus new variables: (1)"fitted", containing the predicted values; and (2) "resid", containing residuals for each observation. The optional PROC PRINT statement simply prints out the first 10 records for verification of what has been saved top the working dataset in memory.

```
          OUTPUT
               OUT=in.mydata
               PREDICTED=fitted
               RESIDUAL=resid;
          PROC PRINT DATA=in.mydata (OBS=10);
```

Note that if no library name is declared and used (here, "in" is the library defining the file folder), then the output datafile (here, mydata, without the "in" prefix) will be saved to memory only and not to file.

SAS supports many other options in addition to PREDICTED under the “Output” statement:

Keyword	Description
COOKD=*names*	Cook’s D influence statistic
COVRATIO=*names*	standard influence of observation on covariance of betas
DFFITS=*names*	standard influence of observation on predicted value
H=*names*	leverage, $x_i(\mathbf{X}'\mathbf{X})^{-1}x_i'$
LCL=*names*	lower bound of a $100(1-\alpha)$% confidence interval for an individual prediction. This includes the variance of the error, as well as the variance of the parameter estimates.
LCLM=*names*	lower bound of a $100(1-\alpha)$% confidence interval for the expected value (mean) of the dependent variable
PREDICTED | P=*names*	predicted values
PRESS=*names*	ith residual divided by $(1-h)$, where h is the leverage, and where the model has been refit without the ith observation
RESIDUAL | R=*names*	residuals, calculated as ACTUAL minus PREDICTED
RSTUDENT=*names*	a studentized residual with the current observation deleted
STDI=*names*	standard error of the individual predicted value
STDP=*names*	standard error of the mean predicted value
STDR=*names*	standard error of the residual
STUDENT=*names*	studentized residuals, which are the residuals divided by their standard errors
UCL=*names*	upper bound of a $100(1-\alpha)$% confidence interval for an individual prediction
UCLM=*names*	upper bound of a $100(1-\alpha)$% confidence interval for the expected value (mean) of the dependent variable

Stata

Following a successful `regress` command, the Stata postestimation `predict` command can be used to create new variables containing the fitted values (here labeled “fitted”) and their standard errors (here labeled “sterror”).

```
predict fitted
predict sterror,stdp
```

After the desired variables are created, from the Stata menu system select File > Save As, to save to file the data with the new variables added.

END OF PREVIEW OF FIRST 30 PAGES OF 410 PAGES

To buy the Kindle version for $9.99, click here.

To buy the entire Statistical Associates “Regression Models” library of 10 statistics books in no-password pdf format on DVD plus one year of free updates for $50, click here.

To buy the entire Statistical Associates library of 50 statistics books in no-password pdf format on DVD plus one year of free updates for $120, click here.

To register for a password-protected pdf version when available, go to http://www.statisticalassociates.com .

Statistical Associates Publishing

Blue Book Series

Association, Measures of
Canonical Correlation
Case Studies
Cluster Analysis
Content Analysis
Correlation
Correlation, Partial
Correspondence Analysis
Cox Regression
Creating Simulated Datasets
Crosstabulation
Curve Estimation & Nonlinear Regression
Delphi Method in Quantitative Research
Discriminant Function Analysis
Ethnographic Research
Evaluation Research
Factor Analysis
Focus Group Research
Game Theory
Generalized Linear Models/Generalized Estimating Equations
GLM (Multivariate), MANOVA, and MANCOVA
GLM (Univariate), ANOVA, and ANCOVA
Grounded Theory
Life Tables & Kaplan-Meier Survival Analysis
Literature Review in Research and Dissertation Writing
Logistic Regression: Binary & Multinomial

Log-linear Models,
Longitudinal Analysis
Missing Values & Data Imputation
Multidimensional Scaling
Multiple Regression
Narrative Analysis
Network Analysis
Neural Network Models
Nonlinear Regression
Ordinal Regression
Parametric Survival Analysis
Partial Correlation
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit and Logit Response Models
Research Design
Scales and Measures
Significance Testing
Social Science Theory in Research and Dissertation Writing
Structural Equation Modeling
Survey Research & Sampling
Testing Statistical Assumptions
Two-Stage Least Squares Regression
Validity & Reliability
Variance Components Analysis
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com