Simple Linear (OLS) Regression

Regression is a method for studying the relationship of a dependent variable and one or more independent variables. Simple Linear Regression tells you the amount of variance accounted for by one variable in predicting another variable.

In this example, we are interested in predicting the frequency of sex among a national sample of adults. The dependent variable is frequency of sex. The independent variables are: age, race, general happiness, church attendance, and marital status.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT sexfreq
  /METHOD=ENTER age Married White happy attend  .
```

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	CHURCH ATTENDANCE, RACE (White =1), GENERAL HAPPINESS, AGE, MARITAL (Married =1)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: FREQUENCY OF SEX DURING LAST YEAR

Model Summary

Model	R	R Square -a-	Adjusted R Square -b-	Std. Error of the Estimate -c-
1	.572[a]	.327	.323	1.651

a. Predictors: (Constant), CHURCH ATTENDANCE, RACE (White =1), GENERAL HAPPINESS, AGE, MARITAL (Married =1)

a. "R Square" = tells us how much of the variance of the dependent variable can be explained by the independent variable(s). Basically, it compares the model with the independent variables to a model without the independent variables. In this case, 33% of the variance is explained by differences in all included variables.

b. "Adjusted R Square" = As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. The Adjusted R Square attempts to produce a more honest value to estimate R Square for the population. In this example, the difference between R Square and Adjusted R Square is minimal.

c. "Std. Error" = The standard error of the estimate (aka, the root mean square error), is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error).

ANOVA[b]

Model		Sum of Squares -f-	df	Mean Square	F -g-	Sig.
1	Regression -d-	1382.662	5	276.532	101.490	.000[a]
	Residual -e-	2850.064	1046	2.725		
	Total	4232.726	1051			

a. Predictors: (Constant), CHURCH ATTENDANCE, RACE (White =1), GENERAL HAPPINESS, AGE, MARITAL (Married =1)

b. Dependent Variable: FREQUENCY OF SEX DURING LAST YEAR

d. The output for "Regression" displays information about the variation accounted for by the model.

e. The output for "Residual" displays information about the variation that is not accounted for by your model. And the output for "Total" is the sum of the information for Regression and Residual.

f. A model with a large regression sum of squares in comparison to the residual sum of squares indicates that the model accounts for most of variation in the dependent variable. Very high residual sum of squares indicate that the model fails to explain a lot of the variation in the dependent variable, and you may want to look for additional factors that help account for a higher proportion of the variation in the dependent variable. In this example, we see that 32.7% of the total sum of squares is made up from the regression sum of squares. You may notice that the R^2 for this model is also .327 (this is not a coincidence!).

g. If the significance value of the F statistic is small (smaller than say 0.05) then the independent variables do a good job explaining the variation in the dependent variable. If the significance value of F is larger than say 0.05 then the independent variables do not explain the variation in the dependent variable. For this example the model does a good job explaining the variation in the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients -i-		Standardized Coefficients -j-	t -k-	Sig. -l-
		B	Std. Error	Beta		
1	(Constant) -h-	5.553	.250		22.191	.000
	AGE	-.055	.003	-.475	-18.255	.000
	MARITAL (Married =1)	1.372	.107	.339	12.810	.000
	RACE (White =1)	-.215	.125	-.045	-1.719	.086
	GENERAL HAPPINESS	-.262	.085	-.081	-3.094	.002
	CHURCH ATTENDANCE	-.067	.020	-.089	-3.436	.001

a. Dependent Variable: FREQUENCY OF SEX DURING LAST YEAR

h. "Constant" represents the Y-intercept, the height of the regression line when it crosses the Y axis. In this example, it is the predicted value of the dependent variable, "frequency of sex," when all other values are zero (0). In this example, the value is 5.553.

i. These are the values for the regression equation for predicting the dependent variable from the independent variable(s). These are *unstandardized* (B) coefficients because they are measured in natural units, and therefore cannot be compared to one another to determine which is more influential.

j. The *standardized* coefficients or betas are an attempt to make the regression coefficients more comparable. Here we can see that the Beta for age has the largest absolute value (-.475), which can be interpreted as having the greatest impact on frequency of sex compared to the other variables in the model. The race variable has the smallest absolute value (-.045) and can be interpreted as having the smallest impact of the dependent variable.

k. The t statistics can help you determine the relative importance of each variable in the model. The t-statistic is calculated by divided the variable's unstandardized coefficient by its standard error. As a guide regarding useful predictors, look for t values well below -1.96 or above +1.96. As you can see, the race variable is the only t-value not within this range (note also that this is the only variable that is not significant).

l. The significance column indicates whether or not a variable is a significant predictor of the dependent variable. The p-value (significance) is the probability that your sample could have been drawn from the population(s) being tested (or that a more improbable sample could be drawn) given the assumption that the null hypothesis is true. A p-value of .05, for example, indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true. As sociologists, we typically look for p-values below .05. For example, in this model the race variable is not significant (p = .086), while the other variables are significant.

Regression Equation

$$SEXFREQ_{predicted} = 5.553 – .055*age – 1.372*marital - .215*race – .262*happy – .067*attend$$

If we plug in values in for the independent variables (age = 35 years; marital = currently married-1; race = white-1; happiness = very happy-1; church attendance = never attends-0), we can predict a value for frequency of sex:

$$SEXFREQ_{predicted} = 5.553- .055*35 – 1.372*1 - .215*1 -.262*1 - .067*0$$
$$= 4.523$$

As this variable is coded, a 35-year old, White, married person with high levels of happiness and who never attends church would be expected to report their frequency of sex between values 4 (weekly) and 5 (2-3 times per week).

If we plug in 70 years, instead, we find that frequency of sex is predicted at 2.531, or approximately 1-2 times per month.

Model Interpretation

Constant = The predicted value of "frequency of sex", when all other variables are 0. Important to note, values of 0 for all variables is not interpretable either (i.e., age cannot equal 0 since in our sample all respondents are between the ages of 18 and 89).

Age = For every unit increase in age (in this case, year), frequency of sex will decrease by .055 units.

Marital Status = For every unit increase in marital status, frequency of sex will decrease by 1.372 units. Since marital status has only two categories, we can conclude that currently married persons have more sex than currently unmarried persons.

Race = For every unit increase in race, frequency of sex will decrease by .215 units. Since race has only two categories, we can conclude that non-White persons have more sex than White persons.

Happiness = For every unit increase in happiness, frequency of sex will decrease by .262 units. Recall that happiness is coded such that higher scores indicate less happiness. For this example, then, higher levels of happiness predict higher frequency of sex.

Church Attendance = For every unit increase in church attendance, frequency of sex decreases by .067 units.

Simple Linear Regression (with nonlinear variables)

It is known that some variables are often non-linear, or curvilinear. Such variables may be age or income. In this example, we include the original age variable and an age squared variable.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT sexfreq
  /METHOD=ENTER age Married White happy attend agesquare  .
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.838	.427		11.329	.000
	AGE	-.020	.017	-.175	-1.190	.234
	MARITAL (Married =1)	1.314	.111	.325	11.883	.000
	RACE (White =1)	-.218	.125	-.046	-1.745	.081
	GENERAL HAPPINESS	-.271	.085	-.084	-3.207	.001
	CHURCH ATTENDANCE	-.067	.020	-.089	-3.423	.001
	AGE-SQUARED	.000	.000	-.303	-2.065	.039

a. Dependent Variable: FREQUENCY OF SEX DURING LAST YEAR

The age squared variable is significant, indicating that age is non-linear.

Simple Linear Regression (with interaction term)

In a linear model, the effect of each independent variable is always the same. However, it could be that the effect of one variable depends on another. In this example, we might expect that the effect of marital status is dependent on gender. In the following example, we include an interaction term, male*married.

To test for two-way interactions (often thought of as a relationship between an independent variable (IV) and dependent variable (DV), moderated by a third variable), first run a regression analysis, including both independent variables (IV and moderator) and their interaction (product) term. It is highly recommended that the independent variable and moderator are standardized before calculation of the product term, although this is not essential.

For this example, two dummy variables were created, for ease of interpretation. Gender was coded such that 1=Male and 0=Female. Marital status was coded such that 1=Currently married and 0=Not currently married. The interaction term is a cross-product of these two dummy variables.

Regression Model (*without* interactions)

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT sexfreq
  /METHOD=ENTER age White happy attend Male Married  .
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.295	.258		20.503	.000
	AGE	-.055	.003	-.469	-18.089	.000
	RACE (White =1)	-.231	.125	-.048	-1.852	.064
	GENERAL HAPPINESS	-.242	.084	-.075	-2.873	.004
	CHURCH ATTENDANCE	-.056	.020	-.074	-2.850	.004
	GENDER (Male =1)	.385	.104	.096	3.709	.000
	MARITAL (Married =1)	1.318	.107	.326	12.262	.000

a. Dependent Variable: FREQUENCY OF SEX DURING LAST YEAR

Regression Model (*with* interactions)

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT sexfreq
  /METHOD=ENTER age White happy attend Male Married Interaction  .
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.136	.262		19.573	.000
	AGE	-.053	.003	-.456	-17.431	.000
	RACE (White =1)	-.263	.125	-.055	-2.109	.035
	GENERAL HAPPINESS	-.251	.084	-.078	-2.986	.003
	CHURCH ATTENDANCE	-.053	.020	-.070	-2.669	.008
	GENDER (Male =1)	.673	.139	.168	4.827	.000
	MARITAL (Married =1)	1.630	.147	.403	11.055	.000
	Male x Married Interaction Term	-.643	.209	-.137	-3.078	.002

a. Dependent Variable: FREQUENCY OF SEX DURING LAST YEAR

The product term should be significant in the regression equation in order for the interaction to be interpretable. This indicates that the effect of being married is significantly different ($p < .05$) for males and females. Marital status has a significant weaker effect for males than females.

Regression Equation

$SEXFREQ_{predicted}$ = 5.136 -.053*age - .263*race -.251*happy - .053*attend + 1.630*married + (.673 - .643*married) * male

Interpretation

Main Effects

The married coefficient represents the main effect for females (the 0 category). The effect for females is then 1.63, or the "marital" coefficient. The effect for males is 1.63 - .643, or .987.

The gender coefficient represents the main effect for unmarried persons (the 0 category). The effect for unmarried is then .673, or the "sex" coefficient. The effect for married is .673 - .643, or .03.

Interaction Effects

For a simple interpretation of the interaction term, plug values into the regression equation above.

Married Men =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**1* + (.673 - . 643**1*) * *1*	= 5.455
Married Women =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**1* + (.673 - . 643**1*) * *0*	= 5.425
Unmarried Men =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**0* + (.673 - . 643**0*) * *1*	= 3.825
Unmarried Women =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**0* + (.673 - . 643**0*) * *0*	= 3.795

In this example (age = 35 years; race = white-1; happiness = very happy-1; church attendance = never attends-0), we can see that (1) for both married and unmarried persons, males are reporting higher frequency of sex than females, and (2) married persons report higher frequency of sex than unmarried persons. The interaction tells us that the gender difference is greater for married persons than for unmarried persons.