Descriptive Statistics

```
PROC MEANS;
VAR age attend happy married racenew male sexfreq;
RUN;
```

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
AGE	AGE OF RESPONDENT	2751	46.2828063	17.3704867	18.0000000	89.0000000
ATTEND	HOW OFTEN R ATTENDS RELIGIOUS SERVICES	2743	3.6624134	2.7020398	0	8.0000000
HAPPY	GENERAL HAPPINESS	1369	1.8210373	0.6289519	1.0000000	3.0000000
MARRIED	Married	2765	0.4589512	0.4984023	0	1.0000000
RACENEW	NEW RACE RECODE	2762	0.7585083	0.4280652	0	1.0000000
MALE	Male	2765	0.4441230	0.4969578	0	1.0000000
SEXFREQ	FREQUENCY OF SEX DURING LAST YEAR	2151	2.8251976	2.0132439	0	6.0000000

"N" = The number of cases with valid responses for each variable. For general happiness, there are 1369 valid responses.

Since we have coded race, gender, and marital status as dummies, the means can be interpreted as the percent of respondents reporting the focus category (1=male, 1=white, 1=married). For, gender, 44% of the respondents are male. For race, 76% of the respondents are white. For marital status, 46% of the respondents are married.

"Mean" = The mean value for all 2751 cases. Respondents have a mean age of 46.28 years.

"Std. Deviation" = The standard deviation is a measure of dispersion around the mean. For example, the mean age for this sample is 46, with a standard deviation of 17. So, 95% of the cases in a normal distribution will be between 29 and 63.

"Minimum" = The minimum possible value for each variable. Recall that values for general happiness are coded (1) very happy, (2) pretty happy, and (3) not too happy.

"Maximum" = The maximum possible value for each variable.

Frequencies

```
PROC FREQ;
TABLES happy / missing;
RUN;
```

In this example, we see that 1396 cases have missing values for the general happiness variable.

The FREQ Procedure

GENERAL HAPPINESS				
HAPPY	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
.	1396	50.49	1396	50.49
1	415	15.01	1811	65.50
2	784	28.35	2595	93.85
3	170	6.15	2765	100.00

The percent column indicates simply the percentage of cases in each of the groups. In this example, 15.01% of the cases have a value of (1) very happy. This column includes only valid values.

Correlations

The correlation tells you the magnitude and direction of the association between two variables.

```
PROC CORR;
VAR happy sexfreq age;
RUN;
```

This cell represents the correlation (and significance and sample size) between age and general happiness. The top value (-.045) is the correlation coefficient. The middle value (.094) is the significance level. The bottom value (1362) is the number of cases. In this example, the correlation is not significant, at the p < .05 level.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	HAPPY	**SEXFREQ**	**AGE**
HAPPY **GENERAL HAPPINESS**	1.00000 1369	-0.10525 0.0006 1060	-0.04534 0.0944 1362
SEXFREQ **FREQUENCY OF SEX DURING LAST YEAR**	-0.10525 0.0006 1060	1.00000 2151	-0.43701 <.0001 2143
AGE **AGE OF RESPONDENT**	-0.04534 0.0944 1362	-0.43701 <.0001 2143	1.00000 2751

Here is the correlation between age of respondent and frequency of sex. The correlation coefficient is -.437 (and is significant). This suggests a negative correlation with moderate magnitude. As age increases, the frequency of sex decreases. The correlation between age and frequency of sex is -.437. If we square this value, we get .190969, or 19.1 out of 100, or 19.1 percent. From this we can claim that 19.1% of the variation in frequency of sex is attributed to respondent's age.

Note
The following general categories indicate a quick way of interpreting correlations.

0.0 – 0.2	Very weak correlation
0.2 – 0.4	Weak correlation
0.4 – 0.7	Moderate correlation
0.7 – 0.9	Strong correlation
0.9 – 1.0	Very strong correlation

T-Test

The Independent Sample T-Test compares the mean scores of two groups on a given variable. In this example, we compare "frequency of sex" for males versus females.

Null Hypothesis: The means of the two groups (males and females) are not significantly different.
Alternate Hypothesis: The means of the two groups (males and females) are significantly different.

```
PROC TTEST;
CLASS sex;
VAR sexfreq;
RUN;
```

The TTEST Procedure

Statistics

Variable	SEX	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
SEXFREQ	1	976	2.9793	3.0994	3.2195	1.831	1.9123	2.0011	0.0612	0	6
SEXFREQ	2	1175	2.4792	2.5974	2.7157	1.9864	2.0667	2.1539	0.0603	0	6
SEXFREQ	Diff (1-2)		0.3322	0.5019	0.6716	1.9401	1.9981	2.0597	0.0865		

This column lists the dependent variable. In this example, it is "frequency of sex."

We can see here that males report higher frequencies of sex than females. However, we cannot tell from here whether or not this difference is significant.

T-Tests					
Variable	**Method**	**Variances**	**DF**	**t Value**	**Pr > |t|**
SEXFREQ	**Pooled**	Equal	2149	5.80	<.0001
SEXFREQ	**Satterthwaite**	Unequal	2124	5.84	<.0001

This column specifies the method for computing standard error of the mean of the difference based on if the assumption of equal variance is used. If we assume equal variance, then we used the “pooled” method. If we do not assume equal variances, then we use the “Satterthwaite” method. When we talk about accounting for both variances, the difference between the two methods is really about how we treat the standard deviations: in the pooled method, we are taking the arithmetic average of the standard deviations and converting this value into a standard error, whereas in the Satterthwaite approximation we are calculating the standard error from the weighted average of the two variances, a subtle, but important difference. The main difference is that the Satterthwaite approximation does not assume equal variances, whereas the pooled method does. In other words, you can *always* use the Satterthwaite method and be correct, but you can only use the pooled method in very specific (and rare) circumstances.

Equality of Variances					
Variable	**Method**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
SEXFREQ	**Folded F**	1174	975	1.17	0.0116

SAS uses an F test for the assumption of equal variance first—this means that the null hypothesis of equal variances is rejected. In other words, variances are unequal, and we use Satterthwaite.

Crosstabs (with Chi-Square)

A crosstabulation displays the number of cases in each category defined by two or more grouping variables.

```
PROC FREQ;
TABLES happy * freqdum / CHISQ;
RUN;
```

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of HAPPY by FREQDUM

HAPPY(GENERAL HAPPINESS)	FREQDUM(SEXFREQ DUMMY) 0	FREQDUM(SEXFREQ DUMMY) 1	Total
1	103 9.72 33.23 22.79	207 19.53 66.77 34.05	310 29.25
2	286 26.98 46.20 63.27	333 31.42 53.80 54.77	619 58.40
3	63 5.94 48.09 13.94	68 6.42 51.91 11.18	131 12.36
Total	452 42.64	608 57.36	1060 100.00

Frequency Missing = 1705

For example, we see that there are 207 cases reporting "very happy" for general happiness and "more than once per month" for frequency of sex.

Statistics for Table of HAPPY by FREQDUM

Statistic	DF	Value	Prob
Chi-Square	2	16.0387	0.0003
Likelihood Ratio Chi-Square	2	16.2971	0.0003
Mantel-Haenszel Chi-Square	1	13.1235	0.0003
Phi Coefficient		0.1230	
Contingency Coefficient		0.1221	
Cramer's V		0.1230	

Effective Sample Size = 1060
Frequency Missing = 1705

WARNING: 62% of the data are missing.

The chi-square measures test the hypothesis that the row and column variables in a crosstabulation are independent. While the chi-square measures may indicate that there is a relationship between two variables, they do not indicate the strength or direction of the relationship.

Note: In this particular dataset, recall that there is 1396 missing cases on the general happiness variable. SAS creates a “warning” when more than 50% of the cases on this test, which may create skewed results.

Chi-Square

The Chi-Square Goodness of Fit Test determines if the observed frequencies are different from what we would expect to find (we expect equal numbers in each group within a variable). Use a Chi-Square Test when you want to know if there is a significant relationship between two categorical variables. In this example, we use “frequency of sex” and “happiness.”

Null Hypothesis: There are approximately equal numbers of cases in each group.
Alternate Hypothesis: There are not equal numbers of cases in each group.

```
PROC FREQ;
TABLES happy / CHISQ;
RUN;
```

The FREQ Procedure

GENERAL HAPPINESS				
HAPPY	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	415	30.31	415	30.31
2	784	57.27	1199	87.58
3	170	12.42	1369	100.00

Frequency Missing = 1396

Chi-Square Test for Equal Proportions	
Chi-Square	418.6866
DF	2
Pr > ChiSq	<.0001

Effective Sample Size = 1369
Frequency Missing = 1396

WARNING: 50% of the data are missing.

We have a Chi-Square value of 418, which is large. Our significance level is .000. We can conclude that there are not equal numbers of cases in each happiness category.

ANOVA

The One-Way ANOVA compares the mean of one or more groups based on one independent variable (or factor). We assume that the dependent variable is normally distributed and that groups have approximately equal variance on the dependent variable.

Null Hypothesis: There are no significant differences between groups' mean scores.
Alternate Hypothesis: There is a significant difference between groups' mean scores.

In this example, we compare "frequency of sex" by church attendance, which was recoded from 9 groups to 3 groups (0=not often, 1=sometimes, 2=often).

```
PROC ANOVA;
CLASS church;
MODEL sexfreq=church;
MEANS church/TUKEY;
MEANS church/BON;
RUN;
```

The ANOVA Procedure

Dependent Variable: SEXFREQ FREQUENCY OF SEX DURING LAST YEAR

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	114.423493	57.211746	14.26	<.0001
Error	2139	8581.451391	4.011899		
Corrected Total	2141	8695.874883			

R-Square	Coeff Var	Root MSE	SEXFREQ Mean
0.013158	70.98556	2.002972	2.821662

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CHURCH	2	114.4234927	57.2117464	14.26	<.0001

$$F = \frac{\text{variance between groups (model)}}{\text{variance expected due to chance (error)}} = \frac{57.212}{4.012} = 14.26$$

If the sample means are clustered closely together (i.e., small differences), the variance will be small; if the means are spread out (i.e., large differences), the variances will be larger.

Our F value is 14.26. Our significance level is .000. We can conclude that there is a significant difference between the three groups. To determine which groups are different from one another, we use the “comparisons” results below.

General Rule: If there are equal number of cases in each group, choose Tukey. If there are not equal numbers of cases of each group, choose Bonferroni. For this example, we will use Bonferroni.

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for SEXFREQ

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	2139
Error Mean Square	4.011899
Critical Value of Studentized Range	3.31681

Comparisons significant at the 0.05 level are indicated by *.**

CHURCH Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 0	0.10709	-0.13871	0.35288	
1 - 2	0.55865	0.29426	0.82303	***
0 - 1	-0.10709	-0.35288	0.13871	
0 - 2	0.45156	0.20724	0.69589	***
2 - 1	-0.55865	-0.82303	-0.29426	***
2 - 0	-0.45156	-0.69589	-0.20724	***

SAS notes a significant difference with an asterisk (*). In this example, we can see that those attending church "often" are significantly different from both of the other groups. However, there is not a significant difference between "not often" and "sometimes."

The ANOVA Procedure

Bonferroni (Dunn) t Tests for SEXFREQ

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	2139
Error Mean Square	4.011899
Critical Value of t	2.39586

Comparisons significant at the 0.05 level are indicated by *.**

CHURCH Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
1 - 0	0.10709	-0.14400	0.35817	
1 - 2	0.55865	0.28857	0.82872	***
0 - 1	-0.10709	-0.35817	0.14400	
0 - 2	0.45156	0.20197	0.70115	***
2 - 1	-0.55865	-0.82872	-0.28857	***
2 - 0	-0.45156	-0.70115	-0.20197	***

SAS notes a significant difference with an asterisk (*). In this example, we can see that those attending church "often" are significantly different from both of the other groups. However, there is not a significant difference between "not often" and "sometimes."

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.

```
PROC REG;
MODEL sexfreq = age racenew happy attend male married;
RUN;
```

The REG Procedure

Dependent Variable: SEXFREQ FREQUENCY OF SEX DURING LAST YEAR

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value[d]	Pr > F
Model[a]	5	1382.66217[c]	276.53243	101.49	<.0001
Error[b]	1046	2850.06406	2.72473		
Corrected Total	1051	4232.72624			

a. The output for Model displays information about the variation accounted for by your model.

b. The output for Error (or Residual) displays information about the variation that is not accounted for by your model. And the output for Corrected Total is the sum of the information for Regression and Error.

c. 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 error 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.

d. 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.

Root MSE	1.65067	R-Square[e]	0.3267
Dependent Mean	2.75475	Adj R-Sq	0.3234
Coeff Var	59.92097		

e. The “R Square” tells us how much of the variance of the dependent variable can be explained by the independent variable(s). In the case of Model 1, 20% of the variance in frequency of sex is explained by differences in age of respondent.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value[g]	Pr > |t|[h]
Intercept[f]	Intercept	1	8.29800	0.29663	27.97	<.0001
AGE	AGE OF RESPONDENT	1	-0.05550	0.00304	-18.26	<.0001
MARITAL	MARITAL STATUS	1	-1.37233	0.10713	-12.81	<.0001
RACENEW	NEW RACE RECODE	1	-0.21538	0.12527	-1.72	0.0859
ATTEND	HOW OFTEN R ATTENDS RELIGIOUS SERVICES	1	-0.06743	0.01962	-3.44	0.0006
HAPPY	GENERAL HAPPINESS	1	-0.26184	0.08464	-3.09	0.0020

f. The “intercept” variable 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 “frequency of sex” when all other values are zero (0).

g. The t statistics can help you determine the relative importance of each variable in the model. As a guide regarding useful predictors, look for t values well below -2 or above +2. For example, for “race” t = 1.72 and not significant.

h. The significance column indicates whether or not a variable is a significant predictor of the dependent variable. For example, the race variable is not significant, while the other variables are significant.

Regression Equation

$SEXFREQ_{predicted}$ = 8.298 – .056*age – 1.372*marital - .215*race – .262*happy – .067*attend

For instance, 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:

$SEXFREQ_{predicted}$ = 8.298 – .056*35 – 1.372*1 - .215*1 – .262*1 – .067*0
= 4.489

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 2.529, or approximately 1-2 times per month.

Model Interpretation

Intercept = The predicted value of "frequency of sex", when all other variables are 0. In this example, a value of 8.298 is not interpretable, since the valid responses for frequency of sex range from 0-6. Important to note, values of 0 for all variables is not interpretable either (i.e., age cannot equal 0).

Age = For every unit increase in age (in this case, year), frequency of sex will decrease by .056 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 .205 units. For example, the difference between non-White (0) to White (1) would be .215 units.

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 non-linear variable)

```
PROC REG;
MODEL sexfreq = age marital racenew happy church agesquar;
RUN;
```

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.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	7.46680	0.49979	14.94	<.0001
AGE	AGE OF RESPONDENT	1	-0.02050	0.01722	-1.19	0.2343
MARITAL	MARITAL STATUS	1	-1.31425	0.11060	-11.88	<.0001
RACENEW	NEW RACE RECODE	1	-0.21831	0.12508	-1.75	0.0812
HAPPY	GENERAL HAPPINESS	1	-0.27140	0.08464	-3.21	0.0014
ATTEND	HOW OFTEN R ATTENDS RELIGIOUS SERVICES	1	-0.06707	0.01959	-3.42	0.0006
AGESQUAR[i]		1	-0.00035233	0.00017065	-2.06	0.0392

i. 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)

```
PROC REG;
MODEL sexfreq = age racenew happy attend male married;
RUN;
```

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	B	7.93136	0.31097	25.51	<.0001
AGE	AGE OF RESPONDENT	1	-0.05477	0.00303	-18.09	<.0001
MARITAL	MARITAL STATUS	B	-1.31795	0.10748	-12.26	<.0001
RACENEW	NEW RACE RECODE	1	-0.23077	0.12458	-1.85	0.0643
HAPPY	GENERAL HAPPINESS	1	-0.24217	0.08430	-2.87	0.0042
ATTEND	HOW OFTEN R ATTENDS RELIGIOUS SERVICES	1	-0.05626	0.01974	-2.85	0.0045
MALE	Male	1	0.38525	0.10387	3.71	0.0002
MARRIED	Married	0	0	.	.	.

Regression Model (with interactions)

```
PROC REG;
MODEL sexfreq = age racenew happy attend male married interact;
RUN;
```

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4.90172	0.24802	19.76	<.0001
AGE	AGE OF RESPONDENT	1	-0.05387	0.00306	-17.63	<.0001
RACENEW	NEW RACE RECODE	1	-0.22534	0.12409	-1.82	0.0697
HAPPY	GENERAL HAPPINESS	1	-0.23012	0.08389	-2.74	0.0062
MALE	Male	1	0.72980	0.13809	5.29	<.0001
MARRIED	Married	1	1.60935	0.14769	10.90	<.0001
INTERACT	Interact	1	-0.67733	0.20924	-3.24	0.0012

The product term should be significant in the regression equation in order for the interaction to be interpretable.

Regression Equation

$SEXFREQ_{predicted}$ = 4.902 -.054*age - .225*race -.230*happy + 1.609*married + (.730 - .677*married) * male

Interpretation

Main Effects

The married coefficient represents the main effect for females (the 0 category). The effect for females is then 1.61, or the "marital" coefficient. The effect for males is 1.61 - .68, or .93.

The gender coefficient represents the main effect for unmarried persons (the 0 category). The effect for unmarried is then .73, or the "sex" coefficient. The effect for married is .73 - .68, or .05.

Interaction Effects

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

Married Men =	$SEXFREQ_{predicted} = 4.902 - .054*age - .225*race - .230*happy + 1.609*\mathit{1} + (.730 - .677*\mathit{1}) * \mathit{1}$	= 4.219
Married Women =	$SEXFREQ_{predicted} = 4.902 - .054*age - .225*race - .230*happy + 1.609*\mathit{1} + (.730 - .677*\mathit{1}) * \mathit{0}$	= 4.166
Unmarried Men =	$SEXFREQ_{predicted} = 4.902 - .054*age - .225*race - .230*happy + 1.609*\mathit{0} + (.730 - .677*\mathit{0}) * \mathit{1}$	= 2.610
Unmarried Women =	$SEXFREQ_{predicted} = 4.902 - .054*age - .225*race - .230*happy + 1.609*\mathit{0} + (.730 - .677*\mathit{0}) * \mathit{0}$	= 2.557

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.

Logistic Regression

Logistic regression is a variation of the regression model. It is used when the dependent response variable is binary in nature. Logistic regression predicts the probability of the dependent response, rather than the value of the response (as in simple linear regression).

In this example, the dependent variable is frequency of sex (less than once per month versus more than once per month).

```
PROC LOGISTIC DESCENDING;
MODEL freqdum = age racenew happy church male married/EXPB;
RUN;
```

SAS chooses the smaller value to estimate its probability. If you include the "descending" option, then SAS will estimate the larger value. If you choose not to include the "descending" option, you will get the same results, except that each B will need to be multipled by negative 1 (-1) and the odds ratios inverted.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1438.921	1146.812
SC	1443.879	1176.563
-2 Log L**a**	1436.921	1134.812

a. "-2 Log L" is used in hypothesis testing for nested models. It is negative two times the log likelihood. Tables typically report the "intercept and covariates" value.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	302.1090	5	<.0001
Score	269.5576	5	<.0001
Wald	203.1289	5	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate**b c**	Standard Error	Chi-Square	Pr > ChiSq	Exp(Est)**d**
Intercept	1	6.7747	0.5080	177.8151	<.0001	875.397
AGE	1	-0.0609	0.00501	148.1741	<.0001	0.941
MARITAL	1	-1.7432	0.1661	110.0822	<.0001	0.175
RACENEW	1	-0.1217	0.1767	0.4745	0.4909	0.885
ATTEND	1	-0.0715	0.0282	6.4130	0.0113	0.931
HAPPY	1	-0.3346	0.1218	7.5451	0.0060	0.716

b. "Estimate" is the estimated coefficient, with the standard error.

c. For this example, the regression equation for the final model is:

$SEXFREQ_{PREDICTED}$ = 6.575 - .061*age - 1.738*marital + .069*race - .070*attend - .334*happiness

d. Recall: When Exp(Est) is less than 1, increasing values of the variable correspond to decreasing odds of the event's occurrence. When Exp(Est) is greater than 1, increasing values of the variable correspond to increasing odds of the event's occurrence.

Intercept = Not interpretable in logistic regression.

Age = Increasing values of age correspond with decreasing odds of having sex more than once a month.

Marital = Increasing values of marital status (married to unmarried) correspond with decreasing odds of having sex more than once a month.

Race = Increasing values of race correspond with increasing odds of having sex more than once a month. Notice that this variable, however, is not significant.

Church Attendance = Increasing values of church attendance correspond with decreasing odds of having sex more than once a month.

Happiness = Increasing values of general happiness correspond with decreasing odds of having sex more than once a month. Recall that happiness is coded such that higher values indicate less happiness.

Odds Ratio Estimates**e**			
Effect	Point Estimate	95% Wald Confidence Limits	
AGE	0.941**f**	0.932	0.950
MARITAL	0.175	0.126	0.242
RACENEW	0.885	0.626	1.252
ATTEND	0.931	0.881	0.984
HAPPY	0.716	0.564	0.909

e. “Exp(Est),” or the odds ratio, is the predicted change in odds for a unit increase in the predictor. When Exp(Est) is less than 1, increasing values of the variable correspond to decreasing odds of the event's occurrence. When Exp(B) is greater than 1, increasing values of the variable correspond to increasing odds of the event's occurrence.

f. If you subtract 1 from the odds ratio and multiply by 100, you get the percent change in odds of the dependent variable having a value of 1. For example, for age:

= 1 - (.941) = .051

= .051 * 100 = 5.1%

The odds ratio for age indicates that every unit increase in age is associated with a 5.1% decrease in the odds of having sex more than once a month.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.9	Somers' D	0.579
Percent Discordant	20.9	Gamma	0.580
Percent Tied	0.2	Tau-a	0.284
Pairs	271051	c	0.790

Logistic Regression (with non-linear variable)

```
PROC LOGISTIC DESCENDING;
MODEL freqdum = age marital racenew happy attend agesquar/EXPB;
RUN;
```

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	4.8068	0.7552	40.5113	<.0001	122.337
AGE	1	0.0289	0.0266	1.1772	0.2779	1.029
MARITAL	1	-1.6731	0.1705	96.3398	<.0001	0.188
RACENEW	1	-0.1216	0.1762	0.4768	0.4899	0.885
ATTEND	1	-0.0713	0.0284	6.2984	0.0121	0.931
HAPPY	1	-0.3529	0.1219	8.3801	0.0038	0.703
AGESQUAR	1	-0.00094	0.000280	11.4037	0.0007	0.999

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

Logistic Regression (with interaction term)

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. Sex was recoded such that 1=Male and 0=Female. Marital status was recoded such that 1=Currently married and 0=Not currently married. The interaction term is a product of these two dummy variables.

Regression Model (without interactions)

```
PROC LOGISTIC DESCENDING;
MODEL freqdum = age racenew happy church male married /EXPB;
RUN;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	3.0471	0.3816	63.7725	<.0001	21.054
AGE	1	-0.0611	0.00506	145.7477	<.0001	0.941
RACENEW	1	-0.1490	0.1782	0.6989	0.4032	0.862
HAPPY	1	-0.3184	0.1228	6.7229	0.0095	0.727
ATTEND	1	-0.0595	0.0286	4.3146	0.0378	0.942
MALE	1	0.4436	0.1483	8.9514	0.0028	1.558
MARRIED	1	1.6983	0.1672	103.1272	<.0001	5.465

Regression Model (with interactions)

```
PROC LOGISTIC DESCENDING;
MODEL freqdum = age racenew happy church male married interact /EXPB;
RUN;
```

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	2.9288	0.3874	57.1643	<.0001	18.704
AGE	1	-0.0601	0.00509	139.1866	<.0001	0.942
RACENEW	1	-0.1729	0.1793	0.9297	0.3349	0.841
HAPPY	1	-0.3223	0.1233	6.8378	0.0089	0.724
ATTEND	1	-0.0563	0.0287	3.8375	0.0501	0.945
MALE	1	0.6488	0.1933	11.2615	0.0008	1.913
MARRIED	1	1.9360	0.2225	75.7215	<.0001	6.931
MALE*MARRIED	1	-0.5036	0.3024	2.7739	0.0958	0.604

The product term should be significant in the regression equation in order for the interaction to be interpretable. In this example, the interaction term is significant at the 0.1 level.

Regression Equation

$FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94*married + (.65 - .50*married) * male

Interpretation

Main Effects

The married coefficient represents the main effect for females (the 0 category). The effect for females is then 1.94, or the "marital" coefficient. The effect for males is 1.94 - .50, or 1.44.

The gender coefficient represents the main effect for unmarried persons (the 0 category). The effect for unmarried is then .65, or the "sex" coefficient. The effect for married is .65 - .50, or .15.

Interaction Effects

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

Married Men =	$FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**1* + (.65 - .50**1*) * *1*	= 2.43
Married Women =	$FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**1* + (.65 - .50**1*) * *0*	= 2.28
Unmarried Men =	$FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**0* + (.65 - .50**0*) * *1*	= 0.49
Unmarried Women =	$FREQDUM_{predicted}$ = 2.93 -.06*age - .17*White -.32*happy - .06*attend + 1.94**0* + (.65 - .50**0*) * *0*	= 0.34

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.

Odds Ratios

Using "married" as the focus variable, we can say that the effect of being married on having sex more than once per month is greater for females.

Females: $e^{1.936}$ = 6.93
Males: $e^{1.432}$ = 4.20

Using "gender" as the focus variable, we can say that the effect of being male on having sex more than once per month is greater for marrieds.

Marrieds: $e^{0.15}$ = 1.16
Unmarrieds: $e^{0.65}$ = 1.92