Introduction to Binary Logistic Regression

Dale Berger
Email: dale.berger@cgu.edu
Website: http://wise.cgu.edu

Page Contents

1607

How does Logistic Regression differ from ordinary linear regression?

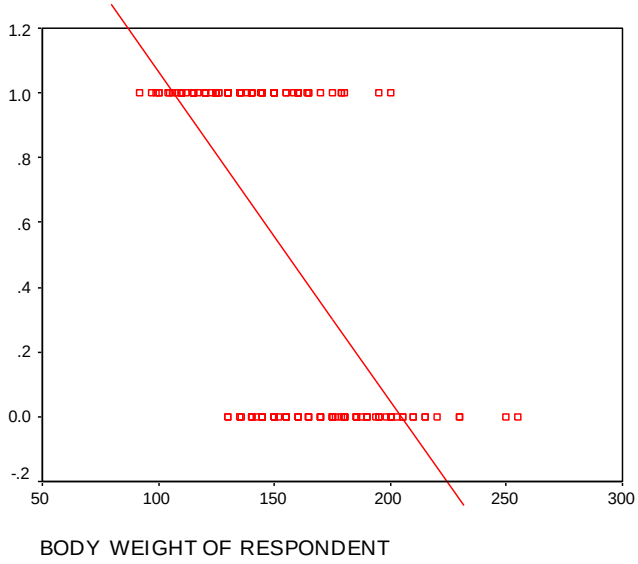
Binary logistic regression is useful where the dependent variable is dichotomous (e.g., succeed/fail, live/die, graduate/dropout, vote for A or B). For example, we may be interested in predicting the likelihood that a new case will be in one of the two outcome categories.

Why not just use ordinary regression? The model for ordinary linear regression (OLS) is

$$Y_i = B_o + B_1X_i + error_i$$

Suppose we are interested in predicting the likelihood that an individual is a female based on body weight. Using real data from 190 Californians who responded to a survey of U.S. licensed drivers (Berger et al., 1990), we could use WEIGHT to predict SEX (coded male = 0, female = 1). An ordinary least squares regression analysis tells us that **Predicted SEX = 2.081 - .01016 * (Body Weight)** and **r = -.649, t(188) = -11.542, p < .001.** A naïve interpretation is that we have a great model.

It is always a good idea to graph data to make sure models are appropriate. A scatter plot gives us intraocular trauma! The linear regression model clearly is not appropriate.



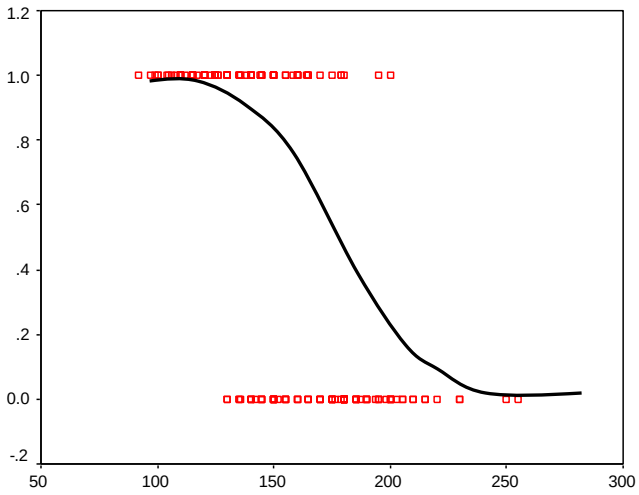
Weight	predicted SEX
100	1.065
150	.557
200	.049
250	-.459

For someone who weighs 150 pounds, the predicted value for SEX is .557. Naively, one might interpret predicted SEX as the probability that the person is a female.

However, the model can give predicted values that exceed 1.000 and or are less than zero, so the predicted values are not probabilities.

The test of statistical significance is based on the assumption that residuals from the regression line are normally distributed with equal variance for all values of the predictor. Clearly, this assumption is violated. The tests of statistical significance provided by the standard OLS analysis are erroneous.

A more appropriate model would produce an estimate of the population average of Y for each value of X. In our example, the population mean approachs SEX=1 for smaller values of WEIGHT and it approaches SEX=0 for larger values of WEIGHT. As shown in the next figure, this plot of means is a curved line. This is what a logistic regression model looks like. It clearly fits the data better than a straight line when the Y variable takes on only two values.



Introduction to the mathematics of logistic regression

Logistic regression forms this model by creating a new dependent variable, the logit(P). If P is the probability of a 1 at for given value of X, the odds of a 1 vs. a 0 at any value for X are P/(1-P). The logit(P) is the natural log of this odds ratio.

Definition : **Logit(P) = ln[P/(1-P)] = ln(odds).** This looks ugly, but it leads to a beautiful model.

In logistic regression, we solve for logit(P) = ***a*** + ***b*** X, where logit(P) is a linear function of X, very much like ordinary regression solving for Y.

With a little algebra, we can solve for P, beginning with the equation $\ln[P/(1-P)] = \boldsymbol{a} + \boldsymbol{b}\, X_i = U_i$. We can raise each side to the power of e, the base of the natural log, 2.71828…

This gives us $P/(1-P) = e^{a + bX}$. Solving for P, we get the following useful equation: $$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Maximum likelihood procedures are used to find the ***a*** and ***b*** coefficients. This equation comes in handy because when we have solved for ***a*** and ***b***, we can compute P.

This equation generates the curved function shown above, predicting P as a function of X. Using this equation, note that as ***a*** + ***b***X approaches negative infinity, the numerator in the formula for P approaches zero, so P approaches zero. When ***a*** + ***b***X approaches positive infinity, P approaches one. Thus, the function is bounded by 0 and 1 which are the limits for P.

Logistic regression also produces a likelihood function [-2 Log Likelihood]. With two hierarchical models, where a variable or set of variables is added to Model 1 to produce Model 2, the contribution of individual variables or sets of variables can be tested in context by finding the difference between the [-2 Log Likelihood] values. This difference is distributed as chi-square with df= (the number of predictors added).

The Wald statistic can be used to test the contribution of individual variables or sets of variables in a model. Wald is distributed according to chi-square.

How well does a model fit?

The most common measure is the Model Chi-square, which can be tested for statistical significance. This is an omnibus test of all of the variables in the model. Note that the chi-square statistic is not a measure of effect size, but rather a test of statistical significance. Larger data sets will generally give larger chi-square statistics and more highly statistically significant findings than smaller data sets from the same population.

A second type of measure is the percent of cases correctly classified. Be aware that this number can easily be misleading. In a case where 90% of the cases are in Group(0), we can easily attain 90% accuracy by classifying everyone into that group. Also, the classification formula is based on the observed data in the sample, and it may not work as well on new data. Finally, classifications depend on what percentage of cases is assumed to be in Group 0 vs. Group 1. Thus, a report of classification accuracy needs to be examined carefully to determine what it means.

A third type of measure of model fit is a pseudo R squared. The goal here is to have a measure similar to R squared in ordinary linear multiple regression. For example, pseudo R squared statistics developed by Cox & Snell and by Nagelkerke range from 0 to 1, but they are not proportion of variance explained.

Limitations

Logistic regression does not require multivariate normal distributions, but it does require random independent sampling, and linearity between X and the logit. The model is likely to be most accurate near the middle of the distributions and less accurate toward the extremes. Although one can estimate P(Y=1) for any combination of values, perhaps not all combinations actually exist in the population.

Models can be distorted if important variables are left out. It is easy to test the contribution of additional variables using hierarchical analyses. However, adding irrelevant variables may dilute the effects of more interesting variables. Multicollinearity will not produce biased estimates, but as in ordinary regression, standard errors for coefficients become larger and the unique contribution of overlapping variables may become very small and hard to detect statistically.

More data is better. Models can be unstable when samples are small. Watch for outliers that can distort relationships. With correlated variables and especially with small samples, some combinations of values may be very sparsely represented. Estimates are unstable and lack power when based on cells with small expected values. Perhaps small categories can be collapsed in a meaningful way. Plot data to assure that the model is appropriate. Are interactions needed? Be careful not to interpret odds ratios as risk ratios.

Comparisons of logistic regression to other analyses

In the following sections we will apply logistic regression to predict a dichotomous outcome variable. For illustration, we will use a single dichotomous predictor, a single continuous predictor, a single categorical predictor, and then apply a full hierarchical binary logistic model with all three types of predictor variables.

We will use data from Berger et al. (1990) to model the probability that a licensed American driver drinks alcoholic beverages (at least one drink in the past year). This data set is available as an SPSS.SAV file called DRIVER.SAV or from Dale.Berger@cgu.edu .

Data Screening

The first step of any data analysis should be to examine the data descriptively. Characteristics of the data may impose limits on the analyses. If we identify anomalies or errors we can make suitable adjustments to the data or to our analyses.

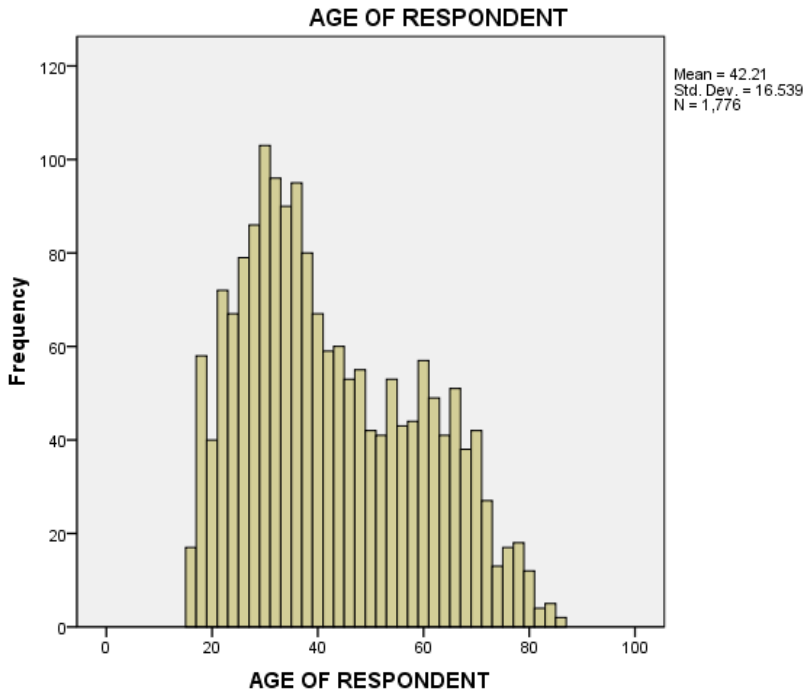
The exercises here will use the variables **age**, **marst** (marital status), **sex2**, and **drink2** (Did you consume any alcoholic beverage in the past year?). We can use SPSS to show descriptive information on these variables.

```
FREQUENCIES VARIABLES=age marst sex2 drink2
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN SKEWNESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM
  /FORMAT=LIMIT(20)
  /ORDER=ANALYSIS.
```

This analysis reveals that we have complete data from 1800 cases for **sex2** and **drink2**, but 17 cases missing data on **age** and 10 cases missing data on **marst**.

To assure that we use the same cases for all analyses, we can filter out those cases with any missing data. Under Data, Select cases, we can select cases that satisfy the condition (age >= 0 & marst >= 0).
Now when we rerun the FREQUENCIES analysis, we find complete data from 1776 on all four variables. We also note that we have reasonably large samples in each subgroup within sex, marital status, and drink2, and age is reasonably normally distributed with no outliers.

We have an even split on sex with 894 males and 882 females. For marital status, there are 328 single, 1205 married or in a stable relationship, 142 divorced or separated, and 101 widowed. Overall, 1122 (63.2%) indicated that they did drink in the past year. Coding for sex2 is male=0 and female=1, and for drink2 none=0 and some=1.



One dichotomous predictor: Chi-square compared to logistic regression

In this demonstration, we will use logistic regression to model the probability that an individual consumed at least one alcoholic beverage in the past year, using sex as the only predictor. In this simple situation, we would probably choose to use crosstabs and chi-square analyses rather than logistic regression. We will begin with a crosstabs analysis to describe our data, and then we will apply the logistic model to see how we can interpret the results of the logistic model in familiar terms taken from the crosstabs analysis. Under Statistics… in Crosstabs, we select Chi-square and Cochran's and Mantel-Haenszel statistics. Under Cells… we select Observed, Column percentages, and both Unstandardized and Standardized residuals. Under Format… select Descending to have the larger number in the top row for the crosstab display.

```
*One dichotomous predictor - first use crosstabs and chi-square.
CROSSTABS
  /TABLES=drink2 BY sex2
  /FORMAT=DVALUE TABLES
  /STATISTICS=CHISQ CMH(1)
  /CELLS=COUNT COLUMN RESID SRESID
  /COUNT ASIS.
```

Crosstabs

drink2 Did you drink last year? * sex2 Sex M=0 F=1 Crosstabulation

			sex2 Sex M=0 F=1		
			0 Male	1 Female	Total
drink2 Did you drink last year?	1 Yes	Count	598	524	1122
		% within sex2 Sex M=0 F=1	66.9%	59.4%	63.2%
		Residual	33.2	-33.2	
		Std. Residual	1.4	-1.4	
	0 No	Count	296	358	654
		% within sex2 Sex M=0 F=1	33.1%	40.6%	36.8%
		Residual	-33.2	33.2	
		Std. Residual	-1.8	1.8	
Total		Count	894	882	1776
		% within sex2 Sex M=0 F=1	100.0%	100.0%	100.0%

Overall, 63.2% of respondents did drink at least one alcoholic beverage in the past year.

We see that the proportion of females who drink is .594 and the proportion of males who drink is .669. The odds that a woman drinks are 524/358 = 1.464, while the odds that a man drinks are 598/296 = 2.020. The odds ratio is 1.464/2.020 = .725. The chi-square test in the next table shows that the difference in drinking proportions is highly statistically significant. Equivalently, the odds ratio of .725 is highly statistically significantly different from 1.000 (which would indicate no sex difference in odds of drinking).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	10.678	1	.001		
Continuity Correction[b]	10.359	1	.001		
Likelihood Ratio	10.690	1	.001		
Fisher's Exact Test				.001	.001
Linear-by-Linear Association	10.672	1	.001		
N of Valid Cases	1776				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 324.79.

b. Computed only for a 2x2 table

Likelihood ratio Chi-square test of independence = 10.690 **(NOT an odds ratio)**

Alternative to Pearson's chi-square approximation

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	10.678	1	.001
Mantel-Haenszel	10.353	1	.001

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.725
ln(Estimate)			-.322
Std. Error of ln(Estimate)			.099
Asymp. Sig. (2-sided)			.001
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	.597
		Upper Bound	.879
	ln(Common Odds Ratio)	Lower Bound	-.516
		Upper Bound	-.129

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Odds Ratio

Calculation of the odds ratio: (524 / 358) / (598 / 296) = .725 or the inverse = 1 / .725 = 1.379
The odds that women drink are .725 times the odds that men drink; The odds that men drink are 1.4 times the odds that women drink. This is not the same as the ratio of probabilities of drinking, where .669 / .594 = 1.13. From this last statistic, we could also say that the probability of drinking is 13% greater for men. However, we could also say that the percentage of men who drink is 7.5% greater than the percentage of women who drink (because 66.9% - 59.4% = 7.5%). **Interpret and present these statistics carefully, attending closely to how they are computed, because these statistics are easily confused.**

One dichotomous predictor in binary logistic regression

Now we will use SPSS binary logistic regression to address the same questions that we addressed with crosstabs and chi-square: Does the variable **sex2** predict whether someone drinks? How strong is the effect?

In SPSS we go to *Analyze, Regression, Binary logistic*… and we select **drink2** as the dependent variable and **sex2** as the covariate. Under Options I selected *Classification Plots*. I selected Paste to save the syntax in a Syntax file. Here is the syntax that SPSS created, followed by selected output.

```
LOGISTIC REGRESSION VAR=drink2
  /METHOD=ENTER sex2
  /CLASSPLOT
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1776	100.0
	Missing Cases	0	.0
	Total	1776	100.0
Unselected Cases		0	.0
Total		1776	100.0

a. If weight is in effect, see classification table for the total number of cases.

Always check the number of cases to verify that you have the correct sample.

Dependent Variable Encoding

Original Value	Internal Value
0 No	0
1 Yes	1

Check the coding to assure that you know which way is up.

sex2 is coded Male = 0, Female = 1

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Did you drink last year?		
			0 No	1 Yes	Percentage Correct
Step 0	Did you drink last year?	0 No	0	654	.0
		1 Yes	0	1122	100.0
	Overall Percentage				63.2

a. Constant is included in the model.

b. The cut value is .500

Because more than 50% of the people in the sample reported that they did drink last year, the best prediction for each case (if we have no additional information) is that the person did drink.

We would be correct 63.2% of the time, because 63.2% (i.e., 1122 of the 1776 people) actually did drink.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.540	.049	120.373	1	.000	1.716

Drinker : Nondrinker ratio
1122 / 654 = 1.716
(odds of drinking)
B=.540; $e^{.540}$ = 1.716

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	sex2	10.678	1	.001
	Overall Statistics		10.678	1	.001

Pearson Chi-square
test of independence
from crosstabs =10.678

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10.690	1	.001
	Block	10.690	1	.001
	Model	10.690	1	.001

Likelihood ratio
Chi-square test of
independence
(NOT an odds ratio)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2326.576[a]	.006	.008

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

For comparison:
Pearson r = -.078, and
R squared = .006

Classification Table[a]

Observed			Predicted		
			Did you drink last year?		
			0 No	1 Yes	Percentage Correct
Step 1	Did you drink last year?	0 No	0	654	.0
		1 Yes	0	1122	100.0
	Overall Percentage				63.2

a. The cut value is .500

[Note: More than half of both males and females drank some alcohol, so the predicted category for everyone, both males and females, is still 'Yes'.]

Wald = $(B/S.E._B)^2$ = $(-.322/.099)^2$ = 10.650
Distributed approximately as chi square with df = 1
Wald for Sex as the only predictor = 10.65.
This value is reported in Table 1.

Odds Ratio:
Odds for females / Odds for males

(524/358) / (598/296)
=1.464 / 2.020 = .725

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sex2	-.322	.099	10.650	1	.001	.725
	Constant	.703	.071	97.916	1	.000	2.020

a. Variable(s) entered on step 1: sex2.

Odds of drinking for
Group 0 (Males)
= 598 / 296 = 2.020

Interpretations: The Wald test for sex is statistically significant, indicating that males are significantly more likely to drink than females (we can look at the data to see the direction of the effect). The odds that a male drinks is 2.020, indicating that there are on average 2.020 males who drink for every male who does not drink. Similarly, there are 1.464 females who drink for every female who does not drink. The ratio of these odds, computed by dividing the odds of drinking for females (Group 1) by the odds of drinking for males (Group 2) is .725. This complex statistic is often misinterpreted.

We can use the logistic regression model to estimate the probability that an individual is in a particular outcome category. In this simple model we have only one predictor, X1 = sex2. We can calculate U = Constant + B1*X1. For a female (X1=1), U_1 = .703 + (-.322)*(1) = .381. We can find an estimate of the probability that Y=1 (i.e., that she drinks) for a female using the following formula:

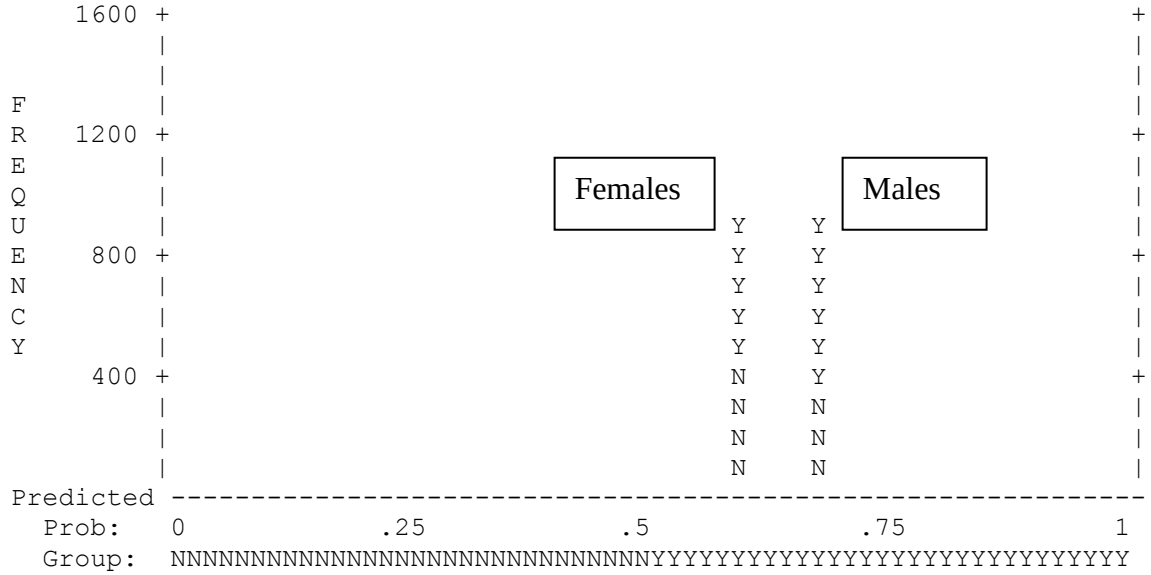
$$\hat{Y}_i = \frac{e^{Ui}}{1+e^{Ui}} = \frac{2.718^{.381}}{1+2.718^{.381}} = \frac{1.46375}{1+1.46375} = .594$$. Check this result in the crosstab table.

Although the logistic model is not very useful for a simple situation where we can simply find the proportion in a crosstab table, it is much more interesting and useful when we have a continuous predictor or multiple predictors.

Below is a plot of predicted p(Y = 1) for every case in our sample, using only sex2 as a predictor. The predicted p(Y = 1) is .594 for females and .669 for males. However, because p(Y = 1) > .50 for all both males and females, we predict Y = 1 for all cases.

Step number: 1

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for Yes
The Cut Value is .50
Symbols: N - No
Y - Yes
Each Symbol Represents 100 Cases.

One continuous predictor: t-test compared to logistic regression

When we have two groups and one reasonably normally distributed continuous variable, we can test for a difference between the group means on the continuous variable with a *t*-test. For illustration, we will compare the results of a standard *t*-test with binary logistic regression with one continuous predictor.

In this example, we will use age to predict whether people drank alcohol in the past year.

```
T-TEST
  GROUPS=drink2(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=age
  /CRITERIA=CIN(.95) .
```

T-Test

Group Statistics

	drink2 Did you drink last year?	N	Mean	Std. Deviation	Std. Error Mean
age AGE OF RESPONDENT	0 No	654	47.78	17.314	.677
	1 Yes	1122	38.96	15.161	.453

On average, people who did not drink at all were 47.78 years old, while those who drank at least some alcohol were 38.96 years old. This implies that the odds of drinking are lower for older people. Logistic regression will give us the actual value.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
age AGE OF RESPONDENT	Equal variances assumed	35.53	.000	11.212	1774	.000	8.818	.787	7.276	10.361
	Equal variances not assumed			10.828	1224.663	.000	8.818	.814	7.221	10.416

The test of statistical significance for age shows t (df = 1774) = 11.212, clear evidence of a relationship between age and drinking, with older people drinking less, on average. In anticipation of logistic analysis where chi-square is used to test the contribution of predictor variables, recall that chi-squared with df = 1 is equal to *z* squared. With df > 1000, *t* is close to *z*, so we can expect that a chi-square test of the age effect will be about $(10.829)^2 = 117$.

One continuous predictor in binary logistic regression

Here we will use SPSS binary logistic regression to address the same questions that we addressed with the *t*-test: Does the variable **age** predict whether someone drinks? If so, how strong is the effect?

In SPSS we go to Analyze, Regression, Binary logistic… and we select **drink2** as the dependent variable and **age** as the covariate. I selected Classification Plots under Options. I clicked Paste to save the syntax in a Syntax file. Here is the syntax that SPSS created, followed by selected output.

```
LOGISTIC REGRESSION VAR=drink2
  /METHOD=ENTER age
  /CLASSPLOT
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1776	100.0
	Missing Cases	0	.0
	Total	1776	100.0
Unselected Cases		0	.0
Total		1776	100.0

a. If weight is in effect, see classification table for the total number of cases.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Did you drink last year?		
			0 No	1 Yes	Percentage Correct
Step 0	Did you drink last year?	0 No	0	654	.0
		1 Yes	0	1122	100.0
	Overall Percentage				63.2

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.540	.049	120.373	1	.000	1.716

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	age	117.524	1	.000
	Overall Statistics		117.524	1	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	118.091	1	.000
	Block	118.091	1	.000
	Model	118.091	1	.000

Recall that the square of the *t* statistic was 117.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2219.176[a]	.064	.088

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

For comparison, Pearson r = -.257, and R squared = .066

Classification Table[a]

Observed			Predicted		
			Did you drink last year?		
			0 No	1 Yes	Percentage Correct
Step 1	Did you drink last year?	0 No	209	445	32.0
		1 Yes	138	984	87.7
	Overall Percentage				67.2

a. The cut value is .500

Wald for age as the only predictor = 111.58. This value is reported in Table 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.033	.003	111.580	1	.000	.968
	Constant	1.963	.146	179.677	1	.000	7.118

a. Variable(s) entered on step 1: age.

Modeled odds of drinking **:** not drinking for someone age 0 is 7.118!

[Model extrapolated beyond observed data!]

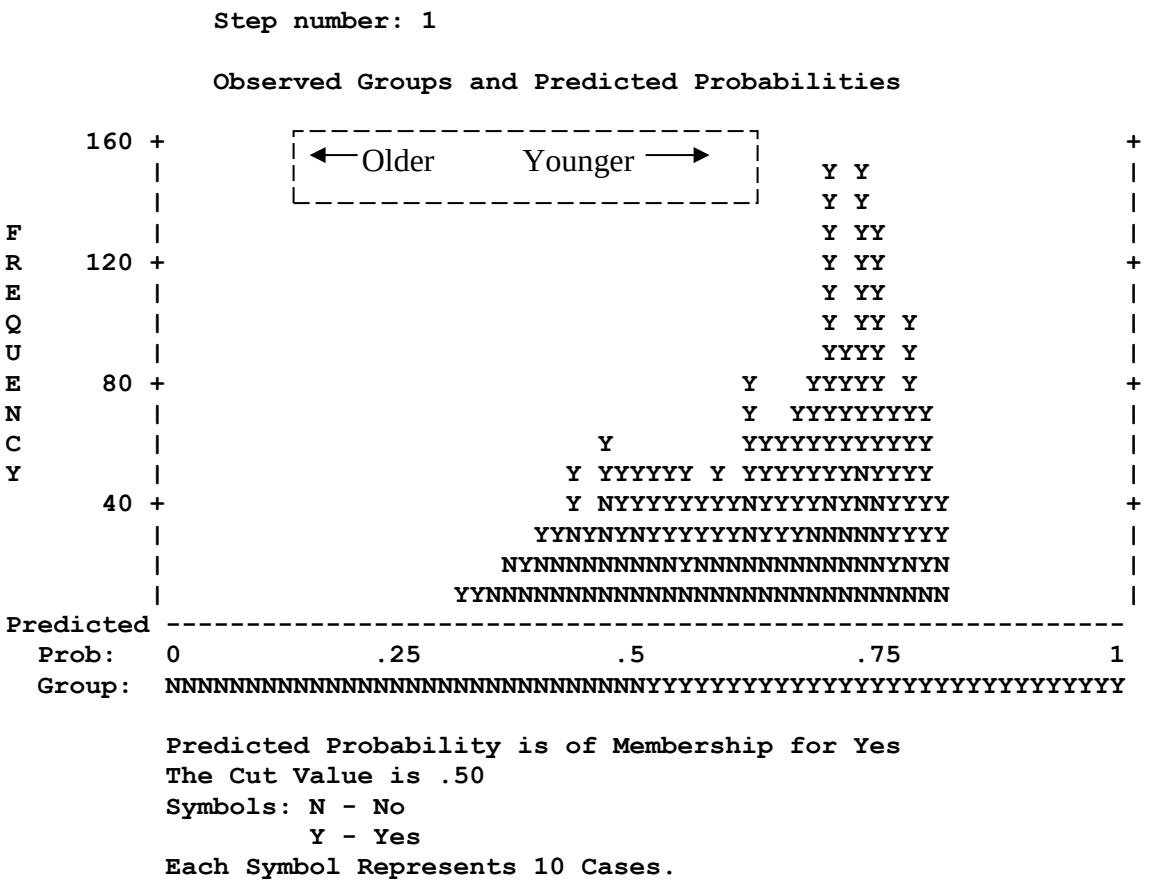
The value of Exp(B) for AGE is .968. This means that the predicted odds of drinking are .968 as great for someone who is one year older than a comparison person. Alternatively, the predicted odds of drinking on average are (1/.968) = 1.033 times greater for a person who is one year younger. This model ignores all other variables and assumes that the log of the odds of drinking is linear with respect to age. We can predict the probability of drinking for an individual at any age.

For someone who is 80 years old, U = 1.963 + (-.033)(80) = -.677.

The modeled probability of drinking is $\hat{Y}_i = \frac{e^{Ui}}{1+e^{Ui}} = \frac{2.718^{-.677}}{1+2.718^{-.677}} = \frac{.508}{1+.508} = .337$

For someone age 21, U = 1.270 and predicted probability of drinking =.781, or 78.1%.

Here is a plot of predicted p(Y = 1) for individuals in our sample. The horizontal axis represents age, with younger people on the right because they have larger predicted probability values (P) compared to older people. This plot is essentially a mirror image of the histogram for age presented earlier.



The default Cut Value is .50. If the estimated P value for an individual is .50 or greater, we predict membership in the Yes group. Alternatively, we could have set the Cut Value at a lower point because the base rate for drinking is 63.2%, clearly higher than 50%. Here we could use 1-.632 = .368. Then, we would predict that any individual with P of .368 or greater would be a drinker.

Consideration of the base rate becomes more important as the base rate deviates farther from .500. Consider a situation where only 1% of the cases fall into the Yes group in the population. We can be correct 99% of the time if we classify everyone into the No group. We might require extremely strong evidence for a case before classifying it as Yes. Setting the Cut Value at .99 would accomplish that goal.

Consideration should be given to the cost and benefits of all four possible outcomes of classification. Is it more costly to predict someone is a drinker when they are not, or to predict someone is not a drinker when they are? Is it more beneficial to classify a drinker as a drinker or to classify a non-drinker as a non-drinker?

The UTIL2 program, available on http://wise.cgu.edu , can help one find the optimal Cut Value to maximize 'utility' of the expected outcome of a classification decision.

One categorical predictor: Chi-square compared to logistic regression

When we have two categorical variables where one is dichotomous, we can test for a relationship between the two variables with chi-square analysis or with binary logistic regression. For illustration, we will compare the results of these two methods of analysis to help us interpret logistic regression.

In this example, we will use marital status to model whether people drank alcohol in the past year.

```
CROSSTABS
  /TABLES=drink2 BY marst
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT COLUMN SRESID.
```

Chi-square with df > 1 is a 'blob' test. Standardized Residuals (SRESID) provides a cell-by-cell test of deviations from the independence model

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
drink2 Did you drink last year? * marst MARITAL STATUS	1776	100.0%	0	.0%	1776	100.0%

drink2 Did you drink last year? * marst MARITAL STATUS Crosstabulation

			marst MARITAL STATUS				
			1 SINGLE	2 MARRIED OR STBL	3 DIV OR SEP	4 WIDOWED	Total
drink2 Did you drink last year?	1 Yes	Count	231	757	93	41	1122
		% within marst MARITAL STATUS	70.4%	62.8%	65.5%	40.6%	63.2%
		Residual	23.8	-4.3	3.3	-22.8	
		Std. Residual	1.7	-.2	.3	-2.9	
	0 No	Count	97	448	49	60	654
		% within marst MARITAL STATUS	29.6%	37.2%	34.5%	59.4%	36.8%
		Residual	-23.8	4.3	-3.3	22.8	
		Std. Residual	-2.2	.2	-.5	3.7	
Total		Count	328	1205	142	101	1776
		% within marst MARITAL STATUS	100.0%	100.0%	100.0%	100.0%	100.0%

The Standardized Residuals can be tested with *z*, so a Std. Residual that exceeds 1.96 in absolute value can be considered statistically significant with two-tailed alpha = .05. These eight tests (one for each cell) are not independent of each other and of course they depend very much on sample size as well as effect size. Here we see that compared to expectations (based on independence of drinking and marital status), significantly fewer single people were non-drinkers, fewer widowed people were drinkers (z = -2.9), and more widowed people were nondrinkers (z = 3.7).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	29.944[a]	3	.000
Likelihood Ratio	29.172	3	.000
Linear-by-Linear Association	21.485	1	.000
N of Valid Cases	1776		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 37.19.

Chi-square with df > 1 is a 'blob' test. The proportion of drinkers is not the same for all marital status groups, but we don't know where the differences exist or how large they are.

The overall tests provided by Pearson Chi-Square and by the Likelihood Ratio Chi-Square indicate that the marital categories differ in the proportions of people who drink. However, these tests are 'blob' tests that don't tell us where the effects are or how large they are. The tests of standardized residuals give us more focused statistical tests, and the proportions who drink in each marital status give us more descriptive measures of the effect.

From the crosstab table, we see that 70.4% of single people drink, while only 40.6% of widowed people drank in the past year. We could also compare each group to the largest group, married, where 62.8% drink. If we were to conduct a series of 2x2 chi-square tests, we would find that single people are significantly more likely to drink than married people, while widowed people are significantly less likely to drink than married people (we could report 2x2 chi square tests to supplement overall statistics).

The odds that a single person drinks is the ratio of those who drink to those who do not drink, 231/97 = 2.381. The odds that a widowed person drinks is 41/60 = .683. Compared to the widowed folks, the odds that single people drink is (231/97) / (41/60) = 2.381/.683 = 3.485 times greater.

Odds ratios are often misinterpreted, especially by social scientists who are not familiar with odds ratios. Here is a Bumble interpretation: "Single people are over three times more likely to drink than widowed people." This is clearly wrong, because 70.4% of single people drink compared to 40.6% for widowed people, which gives 70.4%/40.6% = 1.73, not even twice as likely.

Important lesson: The odds ratio is not the same as a ratio of probabilities for the two groups.

Note the highly significant linear-by-linear association. What does this mean here? Not much – marital status is not even an ordinal variable, much less an interval measure. We just happen to have coded single=1 and widowed=4, so the linear-by-linear association is greater than chance because people with larger scores on marital status are significantly less likely to drink than those with smaller scores, on average.

How else could you describe the difference between single and widowed people in their drinking? The proportion of people who drink in each category would be very useful for description.

Next we will analyze these same data with binary logistic regression and compare the analyses.

One categorical predictor (groups > 2) with binary logistic regression

How is marital status related to whether or not people drink? We will use SPSS binary logistic regression to address this question and compare findings to our earlier 2x4 chi-square analysis.

In SPSS we go to Analyze, Regression, Binary logistic… and we select **drink2** as the dependent variable and **marst** as the covariate. Click Categorical… to define the coding for **marst**. "Indicator" coding is also known as 'dummy' coding, whereby cases that belong in a certain group (e.g., single) are assigned the value of 1 while all others are assigned the value of 0. We need three dummy variables to define membership in all four groups; SPSS creates these for us automatically. There are many other choices for constructing contrasts. However we do it, we need (k – 1) contrasts each with df = 1 to identify membership in k groups. Under Options, I selected Classification Plots. I selected Paste to save the syntax in a Syntax file. Here is the syntax that SPSS created, followed by selected output.

```
LOGISTIC REGRESSION VAR=drink2
  /METHOD=ENTER marst
  /CONTRAST (marst)=Indicator
  /CLASSPLOT
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1776	100.0
	Missing Cases	0	.0
	Total	1776	100.0
Unselected Cases		0	.0
Total		1776	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0 No	0
1 Yes	1

Categorical Variables Codings

			Parameter coding		
		Frequency	(1)	(2)	(3)
marst MARITAL STATUS	1 SINGLE	328	1.000	.000	.000
	2 MARRIED OR STBL	1205	.000	1.000	.000
	3 DIV OR SEP	142	.000	.000	1.000
	4 WIDOWED	101	.000	.000	.000

Check this coding carefully. The first parameter is a dummy code (indicator variable) comparing SINGLE vs. all others combined. Groups 2 and 3 are coded similarly. WIDOWED is the reference group that is omitted from this set of coded variables.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Did you drink last year?		Percentage Correct
			0 No	1 Yes	
Step 0	Did you drink last year?	0 No	0	654	.0
		1 Yes	0	1122	100.0
	Overall Percentage				63.2

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.540	.049	120.373	1	.000	1.716

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	marst	29.944	3	.000
		marst(1)	9.092	1	.003
		marst(2)	.202	1	.653
		marst(3)	.356	1	.551
	Overall Statistics		29.944	3	.000

Pearson χ^2

For variables not in the model, we have tests for the contribution that each single variable would make if it were added to the model by itself. Thus, we see that MARST(1) (single vs. all others combined) would make a statistically significant contribution (p = .003). MARST(2) and MARST(3) would not make significant contributions by themselves. The Score values are chi-square tests of the individual variables; these are reported in Table 1 in the sample write-up.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	29.172	3	.000
	Block	29.172	3	.000
	Model	29.172	3	.000

Likelihood ratio χ^2

Compare to the Likelihood ratio chi-square from CROSSTABS

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2308.094[a]	.016	.022

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

There is no equivalent r or R squared for comparison.

Classification Table[a]

Observed			Predicted		
			Did you drink last year?		
			0 No	1 Yes	Percentage Correct
Step 1	Did you drink last year?	0 No	60	594	9.2
		1 Yes	41	1081	96.3
	Overall Percentage				64.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	marst			28.376	3	.000	
	marst(1)	1.248	.236	27.986	1	.000	3.485
	marst(2)	.905	.211	18.374	1	.000	2.473
	marst(3)	1.022	.269	14.450	1	.000	2.778
	Constant	-.381	.203	3.531	1	.060	.683

a. Variable(s) entered on step 1: marst.

Wald for Marital Status as the only predictor, reported in Table 1.

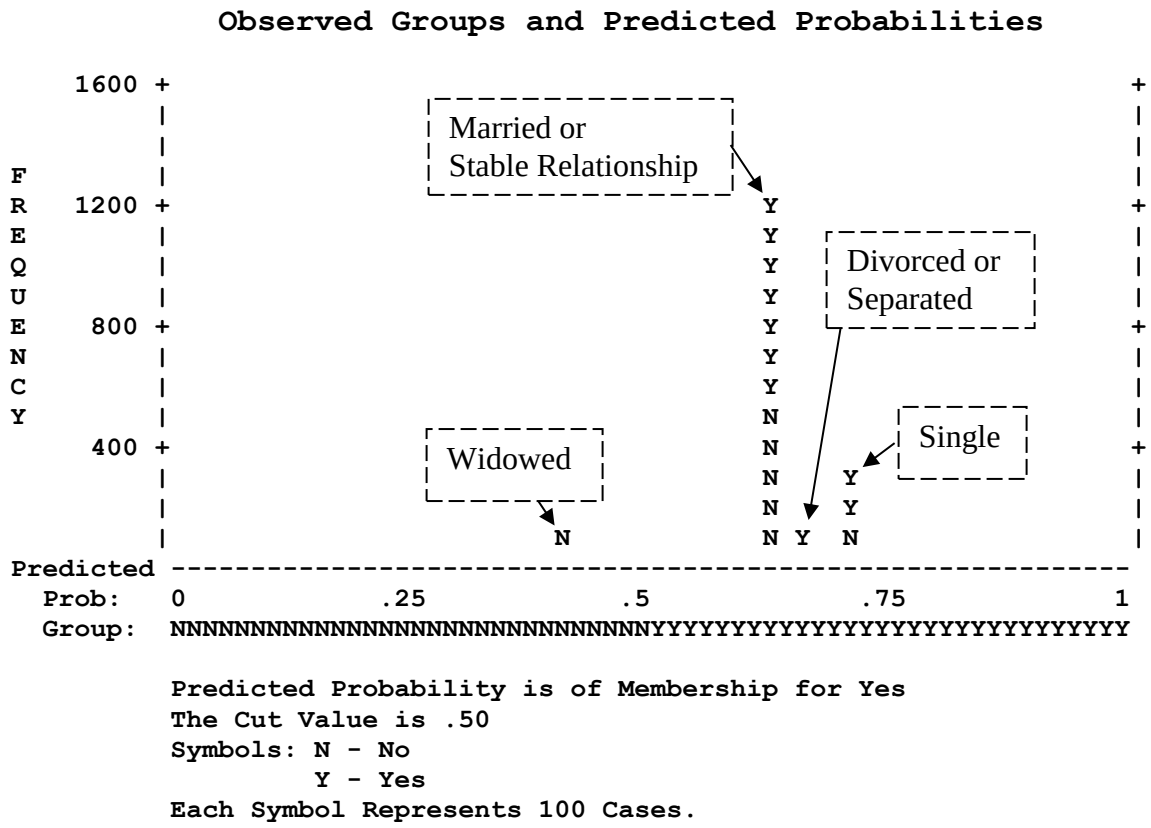
Individual groups are compared to the reference group, Widowed.

When all three indicator variables are in the model together, each test of significance is a test of the unique contribution of one specific variable beyond all of the other variables that are in the model. We may be surprised to see that each of the indicator variables, including MARST(2) and MARST(3) are highly significant ($p < .001$) whereas they were not statistically significant in the table of variables not in the model at Block 0.

The interpretation is a bit tricky the first time you encounter it but it is very important to understand. What is the unique contribution of MARST(3) beyond the other predictors?

The four marital status groups are 1=single, 2=married, 3=divorced, 4=widowed. MARST(1) by itself makes the distinction between 'single' and all the others combined, while MARST(2) by itself makes the distinction between 'married' and all others. Together, those two variables can distinguish each of the first two groups, but they leave 3 and 4 (divorced and widowed) unseparated, both coded zero. MARST(3) can distinguish between 'divorced' and the others. Thus, the unique contribution of MARST(3) in the context of the other dummy variables for marital status is that it can distinguish between 'divorced' and 'widowed.' The test of statistical significance for MARST(3) is a test of the null hypothesis that the odds ratio of drinking / 'not drinking' for divorced vs. widowed is 1.00. If we go back to the crosstab table, we find the odds of drinking for divorced people is 93 / 49 = 1.898, and the odds of drinking for widowed people is 41 / 60 = .683. The odds ratio is 1.898 / .683 = 2.778 = Exp(B) for MARST(3).

Exp(B) for MARST(1) = 3.485. This tells us that compared to the odds of drinking by the reference group (the widowed folks here), the odds that single people drink is 3.485 times greater. This can also be calculated from the frequencies in the crosstab table: (231 / 97) / (41 / 60) = 3.485. Thus, each of the three coded levels of marital status is significantly more likely to drink than the reference group, Widowed. The odds of drinking for the reference group, widowed, is shown as Exp(B) for the constant, .683. (Number of widowed drinkers divided by number of widowed non-drinkers = 41 / 60 = .683).



Here we see that only the Widowed group has a predicted probability of drinking that is less than .500. Because we use the Cut Value of .50, the model will predict all 101 widowed people to be nondrinkers and everyone else to be drinkers.

Look at the Classification Table to see that the 60 + 40 = 101 widowed people were the only people predicted to be nondrinkers.

If we wish to describe the relationship between marital status and drinking, ignoring all other variables, it would be useful to report the percentage of drinkers in each marital group, and perhaps report pairwise tests of statistical significance taken from separate 2x2 crosstab analyses. We would not need logistic regression for that type of analysis.

However, if we wish to control for age and gender when we test for an effect of marital status, logistic regression offers a good option for analysis.

Hierarchical logistic regression with continuous and categorical predictors

Now we can put it all together. We will develop a model to predict the likelihood of drinking based on age, sex, and marital status. When there is logical order such that we are interested in the effects of some variables while controlling for others, we may choose to use a hierarchical model. For example, suppose we consider **age** to be an obvious control variable before we look for sex effects, because women are over-represented among older people. Similarly, if we are interested in marital status for people equivalent on age and sex, we may choose to control for both **age** and **sex** before testing the effects of marital status.

In SPSS, go to Analyze, Regression, Binary Logistic…, select DRINK2 as the dependent measure, select AGE as the first covariate, click Next, select SEX2 as the second covariate, click Next, and select MARST as the third covariate. Now click Categorical, select MARST as a categorical variable. The defaults are OK for the indicator variable, with the last category as the reference group. Click Paste to save the syntax:

```
LOGISTIC REGRESSION VAR=drink2
  /METHOD=ENTER age  /METHOD=ENTER sex2  /METHOD=ENTER marst
  /CONTRAST (marst)=Indicator
  /CLASSPLOT
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Following is selected output from this hierarchical analysis:

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1776	100.0
	Missing Cases	0	.0
	Total	1776	100.0
Unselected Cases		0	.0
Total		1776	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0 No	0
1 Yes	1

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)
MARST MARITAL STATUS	1 SINGLE	328	1.000	.000	.000
	2 MARRIED OR STBL	1205	.000	1.000	.000
	3 DIV OR SEP	142	.000	.000	1.000
	4 WIDOWED	101	.000	.000	.000

Block 0: Beginning Block

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.540	.049	120.373	1	.000	1.716

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	118.091	1	.000
	Block	118.091	1	.000
	Model	118.091	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2219.176	.064	.088

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.033	.003	111.580	1	.000	.968
	Constant	1.963	.146	179.677	1	.000	7.118

a. Variable(s) entered on step 1: age.

Block 2: Method = Enter

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.033	.003	109.496	1	.000	.968
	sex2	-.294	.102	8.263	1	.004	.746
	Constant	2.103	.156	182.070	1	.000	8.188

a. Variable(s) entered on step 1: sex2.

Block 3: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3.127	3	.372
	Block	3.127	3	.372
	Model	129.502	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2207.764	.070	.096

Classification Table[a]

Observed			Predicted		
			Did you drink last year?		Percentage Correct
			No	Yes	
Step 1	Did you drink last year?	No	203	451	31.0
		Yes	140	982	87.5
	Overall Percentage				66.7

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	-.034	.004	89.361	1	.000	.967
	sex2	-.296	.104	8.135	1	.004	.744
	marst			3.135	3	.371	
	marst(1)	-.036	.274	.017	1	.895	.965
	marst(2)	.170	.228	.558	1	.455	1.185
	marst(3)	.277	.284	.948	1	.330	1.319
	Constant	2.013	.321	39.266	1	.000	7.483

a. Variable(s) entered on step 1: marst.

Predicting p(Y=1) for individual cases

What is the estimated probability that a 21-year-old divorced man drinks?
His age is 21, sex2 is 0, marst(1) and marst(2) are 0, and marst(3) is coded as 1

U = (-.034)(21) + (-.296)(0) + (-.036)(0) + (.170)(0) + (.277)(1) + 2.013 = 1.576

$$\hat{Y}_i = \frac{e^{Ui}}{1+e^{Ui}} = \frac{2.718^{1.576}}{1+2.718^{1.576}} = \frac{4.836}{1+4.836} = .829$$

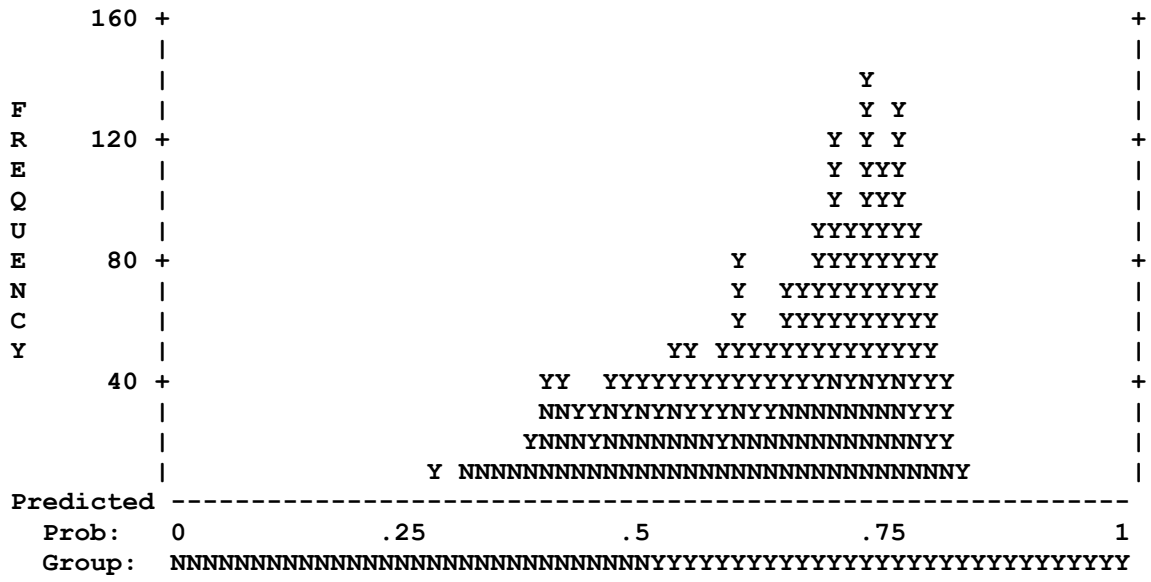
From the model, the probability that a 21-year-old divorced man drinks is 82.9%!

What is the estimated probability that a 90-year-old widowed woman drinks?
We calculate U = (-.034)(90) + (-.296)(1) + 2.013 = -1.315

$$\hat{Y}_i = \frac{e^{Ui}}{1+e^{Ui}} = \frac{2.718^{-1.315}}{1+2.718^{-1.315}} = \frac{.2685}{1+.2685} = .212$$

The probability that a 90-year-old widowed woman drinks is estimated to be 21.2%.

Note that this model does not include any interactions. We could construct interaction terms by multiplying main effects components as we can do with ordinary regression. We should examine the data and model closely to assure ourselves that the model is appropriate to the data. Also, we need to be careful if we extend the model beyond the range of observed data. In this example, the model would predict that a 1-year-old baby is more likely to drink than a 21-year-old!



Our model discriminates quite well between cases in our sample. Predicted probabilities of drinking range from less than 30% to more than 80%.

Data Source:

Berger, D. E., Snortum, J. R., Homel, R. J., Hauge, R., & Loxley, W. (1990). Deterrence and prevention of alcohol-impaired driving in Australia, the United States, and Norway. *Justice Quarterly, 7*(3), 453-465.

Recommended reference:

Tabachnick, B. G. & Fidell, L. S. (2007). Logistic regression. Chapter 10 in *Using multivariate statistics* (5^{th} ed.). Boston, MA: Allyn & Bacon.

Presenting Results

On the next page is a sample write-up of our findings from the series of logistic regression analyses. Note that Table 1 provides three distinct types of information about each predictor. The first chi-square column provides a test of each variable alone, ignoring all other variables. The second chi-square column provides a test of the added contribution of each variable when it is entered, controlling for all prior variables in the hierarchical model. The final model section provides a test of the unique contribution of each variable controlling for all other variables in the final model. This format is only one of many possible ways to present the results.

Results **[Sample Write-up]**

Univariate and hierarchical binary logistic regression models were used to test the contributions of age, sex, and marital status in predicting the likelihood that respondents had consumed any alcoholic beverage in the previous year. As expected, men were more likely than women to report drinking, 66.9% vs. 59.4%, $\chi^2(1, N = 1776) = 10.65$, $p < .001$. People who abstained from alcohol were older on average (M = 47.8 years, SD = 17.3) than people who reported that they had consumed alcohol (M = 39.0, SD = 15.2), $t(1774) = 11.2$, $p < .001$, $d = .54$. Marital status also was related to drinking, with likelihood of drinking 70.4% for single people, 62.8% for those who were married or in a stable relationship, 65.5% for people who were divorced or separated, and 40.6% for widowed people, $\chi^2(3, N = 1776) = 29.94$, $p < .001$. Pairwise comparisons between groups showed that the proportion of drinkers was significantly higher in the first three groups compared to the widowed group (all two-tailed $p < .001$) and that the rate for single people was significantly greater than for married/stable relationship people (two-tailed p=.025). There are six pairwise group comparisons for these four groups, so a conservative Bonferroni adjustment could be applied, with .05/6 = .0083 as the critical p value for statistical significance.

Because of overlap among the predictors (e.g., widowed people are likely to be older women), a research question of interest was whether marital status predicted likelihood of drinking after differences due to age and sex were controlled. Table 1 shows the results of a hierarchical binary logistic regression, where the variables of age, sex, and marital status were entered into the model in that order and the contributions of each variable were tested alone, controlling for previous variables at the point of entry, and controlling for all other variables in the final model.

Age and sex made unique contributions to prediction of drinking in the full model, but marital status did not. Thus, after controlling for sex and age, marital status no longer contributed significantly to predicting drinking, Wald (df=3, N=1776) = 3.14, p=.372.

Table 1

Hierarchical binary logistic regression predicting the likelihood of drinking alcohol in the past year using age, sex, and marital status as predictors (N = 1776)

		Chi-square		Final Model		
Predictor	df	Alone	At entry	B	SE(B)	odds ratio
Age	1	111.58***	111.58***	-.034	.004	.967***
Sex	1	10.65***	8.26**	-.296	.104	.744**
Marital Status	3	28.38***	3.14			
Single	1	9.09**	.02	-.036	.274	.965
Married	1	.20	.56	.170	.228	1.185
Divorced	1	.36	.95	.277	.284	1.319
Constant	1			2.013	.321	7.483***

***$p < .001$, **$p < .01$; Overall model chi-square (5, N = 1790) = 129.50, $p < .001$. When all three marital status variables are in the model, each group is compared to Widowed (the reference group); In the tests for 'Alone,' each group is compared to all other groups combined; Sex is coded Male = 0, Female = 1.

[Note: Reasonable people may choose to present different information in the text or in the table, depending upon their goals. For example, it may be useful to present confidence intervals for the odds ratio.]

How to Graph Logistic Regression Models with Excel

A graph can be an excellent way to show data or a model. Here we demonstrate using the graphing capability of Excel to create a graph showing the predicted probability of drinking as a function of age for single men and women.

First, prepare an Excel file with the exact variables in the same order that they are used in SPSS in the final model "Variables in the equation." In this example, we have the variables Age, Sex, the three dummy variables for marital status, and the constant. Double-click on the table to open the editor, and then you can copy values to paste into Excel. In SPSS there is a blank entry in B for a row headed by MARST, so in Excel you may need to move data to place it into the proper cell. Double-check to make sure you have the correct values in the desired cells under Final B. An advantage of copying values from SPSS rather than entering them by hand into Excel is that you can avoid errors and you have coefficients with many places of precision. Now you can use Excel to compute the P value for a case with any specific characteristics.

In the table below, we pasted the values from SPSS into the column headed Final B. We wish to compute the probability of drinking for single males at ten year intervals from 20 to 90. Because legal drinking age begins at 21, we might choose to begin at 21 rather than 20. We can enter the value 21 for Age, 1 for single, and 1 for the constant because these are the only variables with non-zero values for a 21-year-old single man. The column 'Calc' has a formula (=C7*D7) where C7 refers to the cell in Final B for Age and D7 refers to the cell in Case for age. The result in this example is the product (-.034 * 21 = -70737). The value U is computed as the sum of the values in Calc through the Constant (here the sum U is 1.269192). Now we can compute P (i.e., the predicted value of Y) by using the Excel formula =EXP(E13)/(1+EXP(E13)) where E13 is the cell number for U.

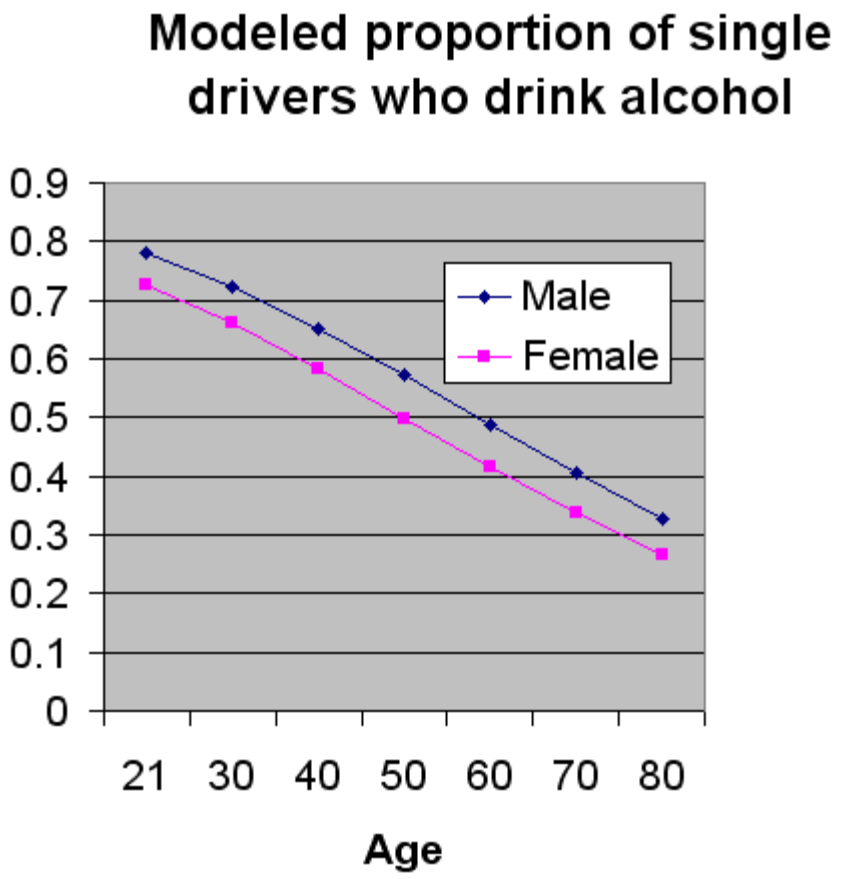
Code	Variable	Final B	Case	Calc
Years	Age	-.034	21	-0.70737
M=0;F=1	Sex	-.296	0	0
Single	Marst(1)	-.036	1	-0.03606
Married	Marst(2)	.170	0	0
Divorced	Marst(3)	.277	0	0
	Constant	2.013	1	2.012627
			U =	1.269192
			P =	0.780604

Age	Male	Female
21	0.780604	0.725748
30	0.724328	0.661502
40	0.652305	0.582526
50	0.57257	0.499077
60	0.488877	0.41568
70	0.405803	0.336847
80	0.327793	0.266155
90	0.258261	0.205697

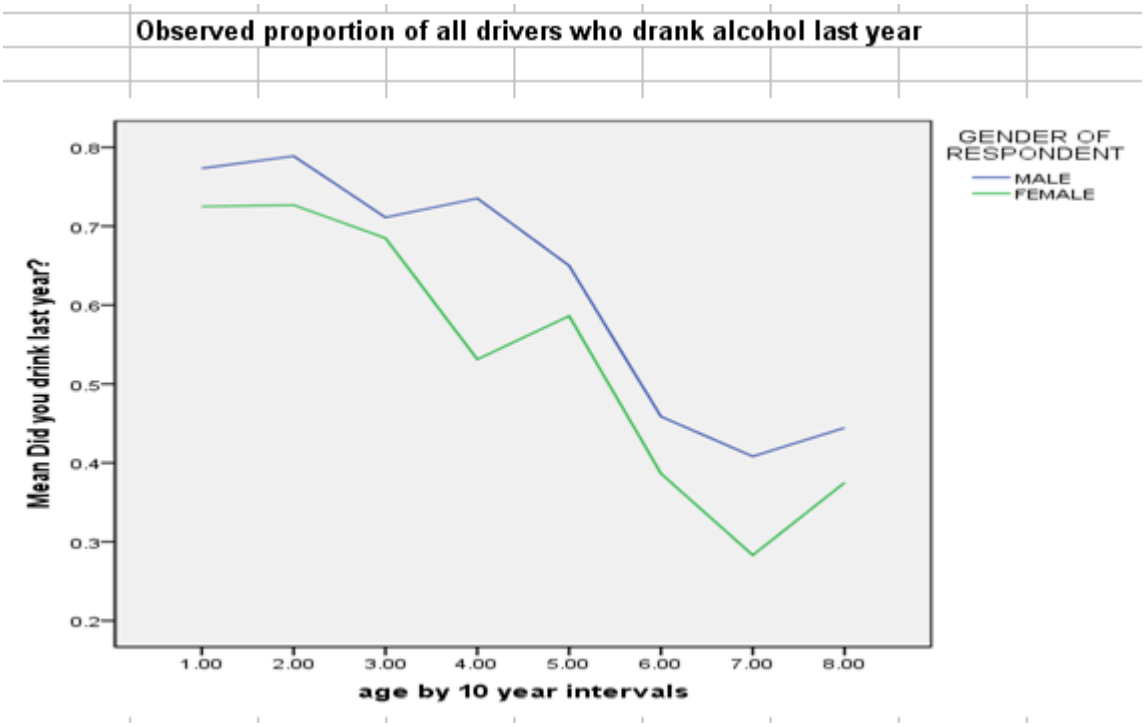
In this example, we find P=.78, the modeled probability that a 21-year-old man drinks alcohol. We can copy this value and paste it into a table that we will use to create the graph (use **Paste Special, value only** because we do not want to paste the formula into the table). The table above on the right was created by methodically changing the age in steps of 10 years for men, and then repeating for women (Sex=1).

When you have completed the table on the right, you can use Excel to create a graph. Highlight the two columns headed by Male and Female (including the headers), and click the Chart Wizard (or click Insert, Chart) to open the Chart Wizard. Select Line graph, click Next to go to Step 2. Click the Series tab, click in the box for Category (X) axis labels, highlight the numbers from 21 through 90 in the data table, click Next to go to Step 3. Enter a title (e.g., Modeled proportion of single drivers who drink alcohol), enter Age for the Category (X) axis, click Next to go to Step 4, and Click Finish. The graph should appear. You can edit the graph within Excel to change colors, markers, labels, etc. An example is on the next page.

Here is an example where I limited the age range between 21 and 80 for the graph (the model included younger cases). This graph could use additional editing to make it more presentable.



With any modeling, it is prudent to verify that the model describes the actual data fairly. As a check for the current example, I recoded age into a new variable with 10-year bins (the first bin is 15-20, and the second bin is 21-29 because legal drinking age begins at 21). In the figure below we see that the youngest drivers (under age 21) were no more likely to drink that the next group (21-29), though the logistic model predicts the highest drinking rate for the youngest drivers. At the top end of the age range we see the real drinking rate has a slight upturn in the actual data. There are very few cases over age 80, so this error may not be consequential. The model looks reasonable for ages 21 to 70. The model does not consider other variables or possible interactions.



How to Graph Logistic Models with SPSS

One can use syntax to generate graphs in SPSS. In this example we will use the coefficients from the final model to generate a graph of modeled proportion of male and female drivers who drink alcohol as a function of age.

We first create a new data file that contains the steps we wish to plot on the X axis (e.g., age 20 to 80 by steps of 5), then provide the equation that uses age (only) to predict probability of drinking, and then create a graph.

* This syntax creates a new data file that has only "Age" in it from 20 to 80 in steps of 5.

```
input program.
  loop #i = 20 to 80 by 5.
  compute age = #i.
  end case.
end loop.
end file.
end input program.
execute.
```

* Enter the B values that you will use from the logistic equation in the compute statement.

```
compute Umale = 2.013 - .036 -.034*age .
compute p_male = exp(Umale)/(1 + exp(Umale)) .
compute Ufemale = 2.013 - .036 -.296 -.034*age .
compute p_female = exp(Ufemale)/(1 + exp(Ufemale)) .
execute.
```

*Create a graph with probability on the Y axis and age on the X axis.

```
GRAPH /LINE(SIMPLE)=VALUE( p_male p_female ) BY age .
```