Logistic Regression

Logistic regression involves a prediction of a binary outcome. Ordinary least squares (OLS) regression assumes a continuous dependent variable *Y* that is distributed approximately normally in the population. Because a binary response variable will not be normally distributed and because the form of the relationship to a binary variable will tend to be nonlinear, we need to consider a different type of model.

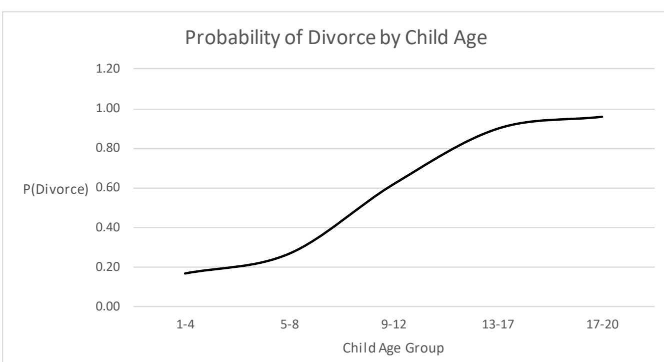
Predicting the Probability that $Y = 1$

For a binary response variable, we can frame the prediction equation in terms of the probability of a discrete event occurring. Usual coding of the response variable is 0 and 1, with the event of interest (e.g., “yes” response, occurrence of an aggressive behavior, or heart attack), so that, if X and Y have a positive linear relationship, the probability that a person will have a score of $Y = 1$ will increase as values of X increase.

For example, we might try to predict whether or not a couple is divorced based on the age of their youngest child. Does the probability of divorce ($Y = 1$) increase as the youngest child’s age (X) increases? If we take a hypothetical example, in which there were 50 couples studied and the children have a range of ages from 0 to 20 years, we could represent this tendency to increase the probability that $Y = 1$ with a graph, grouping child ages into four-year intervals for the purposes of illustration. Assuming codes of 0 and 1 for Y, the average value in each four-year period is the same as the estimated probability of divorce for that age group.

Child Age	Average $E(Y|X)$	Probability of Divorce ($Y = 1$)
1-4	0.17	0.17
5-8	0.27	0.27
9-12	0.62	0.62
13-17	0.90	0.90
17-20	0.96	0.96

The average value within each age group is the expected value for the response at a given value of X, which, with a binary variable, is a conditional probability. Graphing these values, we get



Notice the S-shaped curve. This is typical when we are plotting the average (or expected) values of Y by different values of X whenever there is a positive association between X and Y, assuming a normal and equal distributions for X at each value of Y. As X increases, the probability that $Y = 1$ increases, but not at a consistent rate across values of X. In other words, when children are older, an increasing larger percentage of parents in that child age category divorce, with the increase in divorce probability more dramatic for the middle child age groups.

The Logistic Equation

The S-shaped curve is approximated well by a natural log transformation of the probabilities. In logistic regression, a complex formula is required to convert back and forth from the logistic equation to the OLS-type equation. The logistic equation is stated in terms of the probability that $Y = 1$, which is π, and the probability that $Y = 0$, which is 1 - π.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

The natural log transformation of the probabilities is called the *logit transformation*. The right hand side of the equation, $\alpha + \beta X$, is the familiar equation for the regression line.[1] The left hand side of the equation, the logit $\ln(\pi/(1\text{-}\pi)$, stands in for the predicted value of Y (the observed values are not transformed). So, the predicted regression line is curved line, because of the log function. With estimates of the intercept, α, and the slope β, π can be computed from the equation using the complementary function for the logarithm, e. Given a particular value of X, we can calculate the expected probability that $Y = 1$.

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Because the intercept is the value of Y when X equals 0, the estimate of the probability of $Y = 1$ when $X =$ 0 is $\pi = e^{\alpha}/(1+e^{\alpha})$.[2]

Regression Coefficients and Odds Ratios

Because of the log transformation, our old maxim that β represents "the change in Y with one unit change in X" is no longer applicable. The exponential transformations of the regression coefficient, β, using e^{β} or exp(β) gives us the *odds ratio*, however, which has a more understandable interpretation of the increase in odds for *each unit increase in X*. For illustration purposes, I used grouped ages, in which case, a unit increase would be from one group to the next. Nearly always, we would rather use a more continuous version of age, so a unit increase might be a year. If the odds ratio was 1.22, we would expect approximately a 22% increase in the probability of divorce with each increment in child age. We need to be a little careful about such interpretations, and realize that we are talking about an average percentage increase over all of the range of X. Look back at table of divorce probabilities and the S-shaped figure above. We do not see the same increment in the probability of divorce from the first child age category to the second as we do between the second and the third.

For the special case in which both X and Y are dichotomous, the odds ratio is the probability that Y is 1 when X is 1 compared to the probability that Y is 1 when X is 0, which we have seen before with the analysis of contingency tables. [3]

$$\theta = e^{\beta} = \frac{n_{21}/n_{22}}{n_{11}/n_{12}} = \frac{\pi_{21}/\pi_{22}}{\pi_{11}/\pi_{12}} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$$

Recall that caution is needed in interpreting odds ratios less than 1 (negative relationship) in terms of percentages, because 1/1.22 = .82, where you might be tempted to (incorrectly) interpret the value as indicating an 18% decrease in the probability of divorce instead of more accurately, a 22% decrease.

[1] I follow the notation in the text and use π for the probability even though the observed proportion is being referred to. We also return to using ln for the natural log rather than just "log." The coefficients α and β are unstandardized, not to be confused with the use of b for standardized regression coefficients in ordinary least squares regression.

[2] You may see an alternative but equivalent form of this equation or the equation above used to obtain the proportion from the full model, where 1 is in the numerator: $1/\left(1+e^{-\alpha}\right)$ and $1/\left(1+e^{-(\alpha+\beta x)}\right)$, respectively.

[3] The relative risk can be obtained from the odds ratio, because $RR = OR\left[\left(1-\pi_{1+}\right)/\left(1-\pi_{2+}\right)\right]$ if the marginal frequencies are known.

Odds ratios require some careful interpretation generally because they are essentially in an unstandardized metric. Consider using age as measured by year instead of category in the divorce example. We would expect a smaller percentage increase in the probability that $Y = 1$ for each unit increase in X if X is per year rather per four-year interval increase. If a predictor is measured on a fine-grained scale, such as dollars for annual income, each increment is miniscule and would not the percentage increase in the event to be very large, even if there is a strong magnitude of the relationship between the income and the event. To address this, the X variable is sometimes standardized (*partially standardized* coefficient), to obtain the odds increase for each standard deviation increase in X. Fully standardized coefficients for logistic regression also can be computed, although their meaning is less straightforward than in ordinary least squares regression and there is no universally agreed upon approach. Because software programs do not implement any of them, researchers rarely if ever consider reporting them. A standardized coefficient would have the advantage of interpretation for understanding the relative contribution of each predictor. One can simply calculate the standard deviations of X and Y and standardize the logistic regression coefficient using their ratio as is done in ordinary least squares regression, $\beta^* = \beta_{xy}(s_x/s_y)$. Menard (2010) suggests using the standard deviation of the logit, s^2_{logit}, and the R^2 value as defined for ordinary least squares regression [see the Appendix in Menard (2011) for the computer steps to compute the standardized coefficient].

$$\beta^* = \frac{s_x \beta}{\sqrt{s^2_{\text{logit}} / R^2}}$$

Significance Tests and Confidence Intervals for β and Odds Ratios

The significance of the regression coefficient (that $\beta \neq 0$ in the population) can be tested with the Wald ratio,

$$Wald\ \chi^2 = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}}\right)^2$$

The caret symbol ^ is used by the text to underscore that the coefficient is a sample estimate. The test may be expressed as a z test in some software, where $Wald\ z = \sqrt{Wald\ \chi^2}$. The standard error computation is complex and is derived from the maximum likelihood estimation iterative process.[4] Although the Wald test is the most commonly employed, because it is printed for each coefficient in all software packages, it does not perform optimally in all circumstances. For smaller samples, it tends to be too conservative (i.e., Type II errors are more likely—true relationships are not found to be significant) for large coefficients (Hauck & Donner, 1977; Jennings, 1986).[5] Confidence intervals can also be constructed

$$\hat{\beta} \pm (1.96) s_{\hat{\beta}}$$

where 1.96 is the z critical value for the normal distribution when α = .05 two-tailed. If the confidence interval includes zero, then the coefficient is nonsignificant. Odds ratios may also be presented with confidence limits, in which case, an interval that includes 1.0 is nonsignificant.

Multiple Logistic Regression

Like ordinary least squares regression, a logistic regression model can include two or more predictors. The coefficients and the odds ratios then represent the effect of each independent variable controlling for all of the other independent variable(s) in the model. Each coefficient can be tested for significance, but we may want to also know whether all of the predictors, taken together, account for a significant amount

[4] The standard errors are derived from the information matrix (inverse of the Hessian matrix), computed by the second partial derivatives of the loglikelihood of the matrix of parameter estimates. The Newton-Raphson maximization method is the most common.
[5] Hauck and Donner show that this tendency toward Type II errors increases for more extreme differences between groups (i.e., difference in the proportion of X = 0 and X = 1 groups) and that the Wald sometimes behaves aberrantly for large effects in small samples.

of variance in the dependent variable. Any combination of binary and continuous predictors is possible. For nominal variables with more than two categories, a set of g -1 dummy variables need to be constructed and entered together to capture the differences among the g groups.

Model Fit

Maximum likelihood estimation is used to compute logistic model estimates. The iterative process finds the minimal discrepancy between the observed response, Y, and the predicted response, $\hat{Y}$ (see the handout "Maximum Likelihood Estimation"). The resulting summary measure of this discrepancy is the -2 loglikelihood or -2LL, known as the *deviance* (McCullagh & Nelder, 1989). The larger the deviance, the larger the discrepancy between the observed and expected values. A smaller deviance represents a better fit. The concept is similar to the mean square error (MSE) in ANOVA or regression. Smaller MSE indicates better fit and better prediction. As we add more predictors to the equation, the deviance should get smaller, indicating an improvement in fit. The deviance for the model with one or more predictors is compared to a model without any predictors, called the *null model* or the *constant only* model, which is a model with just the intercept. The now familiar likelihood ratio test is used to compare the deviances of the two models (the null model, L_0 and the full model, L_1).[6]

$$\begin{aligned} G^2 &= Deviance_0 - Deviance_1 \\ &= -2\ln\left(\frac{L_0}{L_1}\right) = \left[-2\log\left(L_0\right)\right] - \left[-2\log\left(L_1\right)\right] \end{aligned}$$

The estimated value of G^2 is distributed as a chi-squared value with df equal to the number of predictors added to the model. The loglikelihoods from any two models can be compared as long as the same number of cases are used and one of the models has a subset of the predictors used in the other model. The special case of the likelihood ratio test in which just one variable is added to the model gives a likelihood ratio test of the significance of a single predictor—the same hypothesis tested by the Wald ratio described above. The Wald test, typically used for testing a single parameter for significance, also can be used to test multiple parameters at once (see Hosmer et al., 2013), although this form of the Wald is not readily available in software packages and seems to be seldomly used for this purpose. A third alternative, the *score* test (or Lagrange multiplier test) is based on partial derivatives of the likelihood function evaluated at β_0 can be used for testing one or more predictors for significance as well. The score test is not always printed or available in software packages (and nearly always just for individual parameters) and is not reported very often by researchers. The Wald, likelihood ratio, and score tests will usually give a very similar result for large sample sizes, and are in fact asymptotically equivalent (Cox & Hinkley, 1972), but the likelihood ratio and score test tend to perform better in many situations (e.g., Hauck & Donner, 1977). The Wald test assumes a symmetric confidence interval whereas the likelihood ratio does not.

Alternative Measures of Fit

Classification Tables. Most regression procedures print a classification table in the output. The classification table is a 2 × 2 table of the observed values on the outcome (e.g., 0="no", 1="yes) and then the values predicted for the outcome by the logistic model. Then the percentage of correctly predicted values (percent of 0s and 1s) correctly predicted by the model is given. Some criteria for deciding what is a correct prediction is need, and by default the program will use the probability that Y = 1 exceeding .5 as "correct." Although authors often report percent correct from the classification as an indicator of fit, it has an inherent problem in the use of .5 as an arbitrary cutoff for correct that is influenced by the base rate value of the probability that Y = 1 (see Box 13.2.8 in Cohen, Cohen, West, & Aiken, 2003). So, I tend not

[6] Important note: G^2 is referred to as "chi-square" in SPSS printouts. And my apologies for the notational whiplash here—I am trying to be faithful to the text's notation. $\ln$ is the natural log, so $\ln = \log$ in this context. A special case of this equation is the same as the G^2 equation we examined in connection with the 2 × 2 contingency table, G^2 is a function of the observed (n_{ij}) and expected frequencies (μ_{ij}) across each of the cells.

$$G^2 = 2\sum_i^I \sum_j^J n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right)$$

to use the percent correctly classified and tend to take it with a grain of salt when other researchers report it.

Hosmer-Lemeshow Test. The likelihood ratio test (G^2) does not always perform well (Hosmer & Lemeshow, 1980; McCullagh 1985; Xu, 1996), especially when data are *sparse*. The term "sparse" refers to a circumstance in which there are few observed values (and therefore few expected values) in the cells formed by crossing all of the values of all of the predictors. An alternative test developed by Hosmer and Lemeshow (1980) is commonly printed with logistic regression output. The Hosmer-Lemeshow test is performed by dividing the predicted probabilities into deciles (10 groups based on percentile ranks) and then computing a Pearson chi-square that compares the predicted to the observed frequencies (in a 2 × 10 table). Lower values (and nonsignificance) indicate a good fit to the data and, therefore, good overall model fit. Unfortunately, even Hosmer and Lemeshow (2013) do not recommend using their test unless the sample size is at least 400 (when sparseness may not be as much of a problem) because of insufficient power; and it has other potential problems (Allison, 2014; Hosmer, Hosmer, Le Cessie, & Lemshow, 1997). There are several other potential alternative fit tests, such as the standardized Pearson test or the Stukel test, which are not widely available in software packages and appear to be less often used by researchers (see Allison, 2014 for an excellent summary), some of which may also require larger sample sizes for sufficient power (Hosmer et al., 2013).

Information Criteria. You will also hear about several absolute fit indices, such as the Akaike information criteria (AIC) or Bayesian information criteria (BIC), which can be useful for comparing models (lower values indicate better fit). (SPSS does not print several other global fit indices that are sometimes used by researchers testing logistic regression models). The AIC and BIC do not have values that are informative by themselves because they are fairly simply derived from the deviance using adjustments for sample size and number of predictors. Because the deviance itself depends on the size of the model, variances of the variables involved, and other factors, it has no possible standard of magnitude and thus neither does the AIC or BIC (there are no statistical tests for these indices and no cutoff for what constitutes a good fit). Indices like the AIC and BIC are occasionally used, however, to try to compare non-nested models (models that do not have the same cases and where one model has a subset of predictors from the other model). When models are nested, the likelihood ratio (difference in deviances) can be used as a statistical test (chi-square value), so there is not really a need for the AIC or BIC in that case. The AIC and BIC are perhaps the most commonly used but there are several other similar indices, such as the AICC and aBIC. The equations below show the AIC and BIC are fairly simply derived of the deviance (-2LL value), shown below with p as the number of predictors and n as the sample size.

$$AIC = -2LL + 2(p+1)$$
$$BIC = -2LL + \log(n)(p+1)$$

R^2 for Logistic Regression. In logistic regression, there is no true R^2 value as there is in OLS regression. However, because deviance can be thought of as a measure of how poorly the model fits (i.e., lack of fit between observed and predicted values), an analogy can be made to sum of squares residual in ordinary least squares. The proportion of *unaccounted* for variance that is reduced by adding variables to the model is the same as the proportion of variance accounted for, or R^2.

$$R^2_{\text{logistic}} = \frac{-2LL_{null} - 2LL_k}{-2LL_{null}} \qquad R^2_{OLS} = \frac{SS_{total} - SS_{residual}}{SS_{total}} = \frac{SS_{regression}}{SS_{total}}$$

Where the null model is the logistic model with just the constant and the k model contains all the predictors in the model.

There are a number of pseudo-R^2 values that have been proposed using this general logic, including the Cox and Snell (Cox & Snell, 1989; Cragg & Uhler, 1970; Maddala,1983), Nagelkerke (1991), McFadden

(1974), and Tjur (2009) indexes, among others (see Allison, 2014, for a review). As two common examples, consider the following:

Cox & Snell Pseudo-R^2

$$R^2 = 1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{2/n}$$

Because the Cox and Snell R-squared value cannot reach 1.0, Nagelkerke modified it. The correction increases the Cox and Snell version to make 1.0 a possible value for R-squared.

Nagelkerke Pseudo-R^2

$$R^2 = \frac{1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{2/n}}{1 - \left(-2LL_{null}\right)^{2/n}}$$

At this point, there does not seem to be much agreement on which R-square approach is best (see https://statisticalhorizons.com/r2logistic for a brief discussion and references), and researchers do not seem to report any one of them as often as they should. My recommendation for any that you choose to use, do not use them as definitive or exact values for the percentage of variance accounted for and to make some reference to the "approximate percentage of variance accounted for".

Logistic Regression, Chi-squared, and Loglinear Models Compared

As you might have wondered by now, the simple logistic regression model with a binary independent variable could be used to analyze a two-way contingency table. And, in fact, for that special case, the likelihood ratio test from the contingency table analysis and the logistic regression are the same. Though the loglinear model does not distinguish between explanatory and response variables—all are essentially treated as response variables—the simple logistic and the likelihood ratio test from the loglinear model in the 2 × 2 case will equal the likelihood ratio test from the logistic regression. So, in the simple case, these analyses converge. With more complex analyses, it becomes more difficult to always see the connection. The three-way contingency table analysis also relates to the logistic regression model. A logistic model that tests the same hypothesis as tests from the loglinear and three-way contingency tests can be constructed if we consider a logistic model with more than one predictor (e.g., X and Z predicting Y).

Software Examples

The Quinnipiac polling data[7] is reanalyzed with simple logistic with a binary predictor. Compare these results to the results from the contingency table analyses in the "Analysis of Contingency Tables" and the loglinear analyses from the "Loglinear Models" handouts.

SPSS

```
logistic regression vars=response with ind
   /print=summary ci(95) goodfit iter(1).  *note: CI value must be a whole number.
```

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	982	99.4
	Missing Cases	6	.6
	Total	988	100.0
Unselected Cases		0	.0
Total		988	100.0

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

[7] Data source: https://poll.qu.edu/georgia/release-detail?ReleaseID=3679. Note that the data extrapolated cell sample sizes and used some rounding, so the results should be taken as only approximate.

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	1358.146	.114
	2	1358.146	.114

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 1358.146

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients	
			Constant	ind vs affil
Step 1	1	1357.022	.071	.149
	2	1357.022	.071	.150

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1358.146

d. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.123	1	.289
	Block	1.123	1	.289
	Model	1.123	1	.289

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1357.022[a]	.001	.002

a. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

Contingency Table for Hosmer and Lemeshow Test

		response intended vote = 0 Trump		response intended vote = 1 Biden		
		Observed	Expected	Observed	Expected	Total
Step 1	1	338	338.000	363	363.000	701
	2	125	125.000	156	156.000	281

Classification Table[a]

Observed			Predicted		
			response intended vote		Percentage Correct
			0 Trump	1 Biden	
Step 1	response intended vote	0 Trump	0	463	.0
		1 Biden	0	519	100.0
	Overall Percentage				52.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ind vs affil	.150	.142	1.121	1	.290	1.162	.880	1.535
	Constant	.071	.076	.891	1	.345	1.074		

a. Variable(s) entered on step 1: ind vs affil.

Note the corrected confidence limits above, which now include 1.0.

R

```
> logmod <- glm(response ~ ind, data = d, family = "binomial")
> summary(logmod)

Call:
glm(formula = response ~ ind, family = "binomial", data = d)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.273   -1.208    1.085    1.147    1.147

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.07136    0.07559   0.944    0.345
ind          0.15019    0.14186   1.059    0.290

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1358.1  on 981  degrees of freedom
Residual deviance: 1357.0  on 980  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 1361

Number of Fisher Scoring iterations: 3

> #easy way to get odds ratios
> exp(cbind(OR=coef(logmod), confint(logmod)))
Waiting for profiling to be done...
                  OR     2.5 %   97.5 %
(Intercept) 1.073964 0.9261414 1.245707
ind         1.162050 0.8804390 1.535866

> #obtain psuedo-R-sq values with modEvA package
> library(modEvA)
> RsqGLM(model=logmod)
$CoxSnell
[1] 0.001143395

$Nagelkerke
[1] 0.001526185

$McFadden
[1] 0.0008271988

$Tjur
[1] 0.001142238

$sqPearson
[1] 0.001142238
```

SAS

```
proc logistic data=one order=data descending; ;
model response=ind;
run;
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.ONE	
Response Variable	response	intended vote
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	988
Number of Observations Used	982

Response Profile

Ordered Value	response	Total Frequency
1	Trump	463
2	Biden	519

Probability modeled is response='Trump'.

NOTE: 6 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1360.146	1361.022
SC	1365.035	1370.802
-2 Log L	1358.146	1357.022

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.1235	1	0.2892
Score	1.1217	1	0.2896
Wald	1.1209	1	0.2897

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0714	0.0756	0.8912	0.3452
ind	1	-0.1502	0.1419	1.1209	0.2897

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
ind	0.861	0.652	1.136

```
Association of Predicted Probabilities and Observed Responses

Percent Concordant       21.9    Somers' D    0.031
Percent Discordant       18.9    Gamma        0.075
Percent Tied             59.2    Tau-a        0.015
Pairs                  240297    c            0.515
```

Multiple Logistic Examples

To illustrate multiple logistic regression, I used data from the Late Life Study of Social Exchanges (LLSSE; Sorkin & Rook, 2004) to predict self-reported heart disease. Predictors include sex (`w1sex`), vigorous physical activity (`w1active`), depression symptomatology from the brief 9-item version (Santor & Coyne, 1997) of the Center for Epidemiologic Studies-Depression scale (Radloff, 1977), and a measure of negative social exchanges (`w1neg`; Newsom, Rook, Nishishiba, Sorkin, & Mahan), which assesses the frequency of interpersonal conflicts.

SPSS

```
logistic regression vars=w1hheart with w1sex w1activ w1cesd9 w1neg
   /print=summary ci(95) goodfit iter(1).   *note: CI value must be a whole number.
```

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	602.480 a	.033	.055

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23.238	4	.000
	Block	23.238	4	.000
	Model	23.238	4	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	A1-sex of R	-.978	.214	20.824	1	.000	.376	.247	.572
	N11-how often vigerous phy activities (Health)	-.041	.048	.721	1	.396	.960	.874	1.055
	9-item CES-D	.035	.022	2.550	1	.110	1.036	.992	1.082
	negative exchanges-total (mean)	.068	.186	.132	1	.716	1.070	.743	1.542
	Constant	-1.199	.206	33.922	1	.000	.301		

a. Variable(s) entered on step 1: A1-sex of R, N11-how often vigerous phy activities (Health), 9-item CES-D, negative exchanges-total (mean).

Correct confidence limits now in the above table.

R

```
> logmod <- glm(w1hheart ~ w1sex + w1activ + w1cesd9 + w1neg, data = d, family = "binomial")
> summary(logmod)


Call:
glm(formula = w1hheart ~ w1sex + w1activ + w1cesd9 + w1neg, family = "binomial",
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0471  -0.6846  -0.4980  -0.4545   2.2390

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
```

```
(Intercept) -1.19911    0.20588  -5.824 0.00000000574
w1sex       -0.97841    0.21440  -4.563 0.00000503352
w1activ     -0.04065    0.04788  -0.849         0.396
w1cesd9      0.03539    0.02216   1.597         0.110
w1neg        0.06780    0.18643   0.364         0.716

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.72  on 691  degrees of freedom
Residual deviance: 602.48  on 687  degrees of freedom
  (32 observations deleted due to missingness)
AIC: 612.48

Number of Fisher Scoring iterations: 4

> #obtain odds ratios
> exp(cbind(OR=coef(logmod), confint(logmod)))
Waiting for profiling to be done...
                   OR     2.5 %    97.5 %
(Intercept) 0.3014610 0.1998632 0.4485558
w1sex       0.3759076 0.2458771 0.5707173
w1activ     0.9601616 0.8724972 1.0530681
w1cesd9     1.0360220 0.9909311 1.0812665
w1neg       1.0701472 0.7324376 1.5282734


> #obtain psuedo-R-sq values with modEvA package
> library(modEvA)
> RsqGLM(model=logmod)


$CoxSnell
[1] 0.03302352

$Nagelkerke
[1] 0.05548855

$McFadden
[1] 0.03713834

$Tjur
[1] 0.03423869

$sqPearson
[1] 0.03373417
```

SAS

```
proc logistic data=one order=data descending; ;
model w1hheart=w1sex w1activ w1cesd9 w1neg;
run;
```

```
                     Number of Observations Read         724
                     Number of Observations Used         692

                                  Response Profile

                          Ordered                          Total
                            Value      w1hheart        Frequency

                                1      yes                   116
                                2      no                    576

                        Probability modeled is w1hheart='yes'.

NOTE: 32 observations were deleted due to missing values for the response or explanatory variables.

                              Model Convergence Status

                   Convergence criterion (GCONV=1E-8) satisfied.

                                Model Fit Statistics

                                                      Intercept
```

Criterion	Intercept Only	Intercept and Covariates
AIC	627.718	612.480
SC	632.258	635.178
-2 Log L	625.718	602.480

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	23.2381	4	0.0001
Score	23.6844	4	<.0001
Wald	22.6227	4	0.0002

The LOGISTIC Procedure
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1991	0.2059	33.9219	<.0001
w1sex	1	-0.9784	0.2144	20.8240	<.0001
w1activ	1	-0.0407	0.0479	0.7209	0.3958
w1cesd9	1	0.0354	0.0222	2.5497	0.1103
w1neg	1	0.0678	0.1864	0.1322	0.7161

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
w1sex	0.376	0.247	0.572
w1activ	0.960	0.874	1.055
w1cesd9	1.036	0.992	1.082
w1neg	1.070	0.743	1.542

Association of Predicted Probabilities and Observed Responses

Percent Concordant	64.4	Somers' D	0.291
Percent Discordant	35.3	Gamma	0.292
Percent Tied	0.3	Tau-a	0.081
Pairs	66816	c	0.645

Sample Write-Up

To identify factors that predict self-reported heart disease in a sample of older adults, a multiple logistic regression analysis was conducted, simultaneously entering sex, self-reported physical activity, depression scores, and negative social exchanges into the model. The results indicated that, together, the predictors accounted for a significant amount of variance in success, likelihood ratio $\chi^2(4) = 23.238$, $p < .001$. The Nagelkerke pseudo-R^2 indicated approximately 6% of the variance in heart disease was accounted for by the predictors overall. Out of all of the predictors in the model, only sex was a significant independent predictor of heart disease, $b = -.978$, $SE = .214$, p <.001, with women more than two and a half times less likely to report heart disease, *OR* = .376 (where the odds for men vs. women = 1/.376 = 2.660) after controlling for activity level, depression, and negative social exchanges.[8]

References and Further Reading

Allison, P. D. (2014). Measures of fit for logistic regression. *SAS Global Forum, Washington, DC*.
Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*. Second Edition. Chapman & Hall.
Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
Cragg, J.G. & Uhler, R.S. (1970). "The demand for automobiles." *The Canadian Journal of Economics, 3*: 386-406.
Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, *81*(394), 471-476.
Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association*, *72*(360a), 851-853.
Hosmer D.W. and S. Lemeshow (1980). A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics A10:1043-1069.
Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. New York: Wiley.
Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine, 16*, 965-980.
Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, *81*(394), 471-476.
Long, J.S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.
Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistics in linear exponential family models. *International Statistical Review 53*, 61–67.
McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics* (pp. 105-142). Academic Press.
Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications, second edition*. Sage Publications.
Menard, S. (2011). Standards for standardized logistic regression coefficients. *Social Forces, 89*, 1409-1428.
Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika,* 78, 691-692.
Newsom, J.T., Rook, K.S., Nishishiba, M., Sorkin, D., & Mahan, T.L. (2005). Understanding the relative importance of positive and negative social exchanges: Examining specific domains and appraisals. *Journals of Gerontology: Psychological Sciences, 60B*, P304-P312.
Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychology and Measurement, 1, 385-401.
Santor, D. A. & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9*, 233–243.
Sorkin, D. H., & Rook, K. S. (2004). Interpersonal control strivings and vulnerability to negative social exchanges in later life. *Psychology and Aging, 19*, 555–564.
Tjur, T. (2009) Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician, 63*, 366-372.

[8] Had negative social exchanges been significant, we might say that the odds of heart disease increased by about 7% for each unit increase on the scale, OR = 1.070. Depending on the number of predictors and there is a table, non-significant coefficients might be reported in the text.