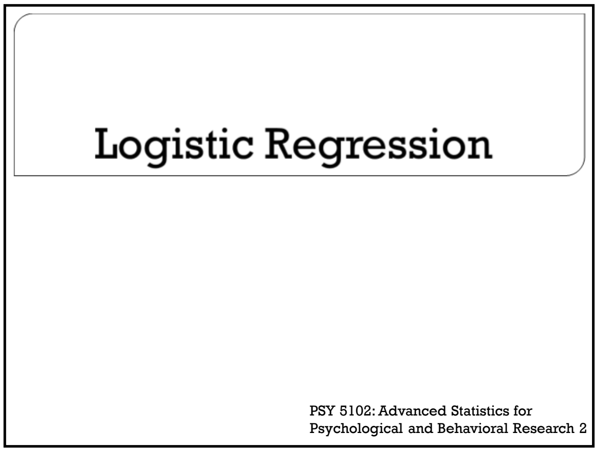
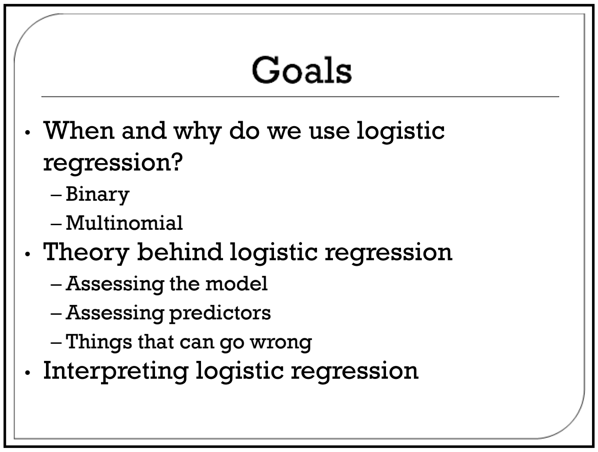


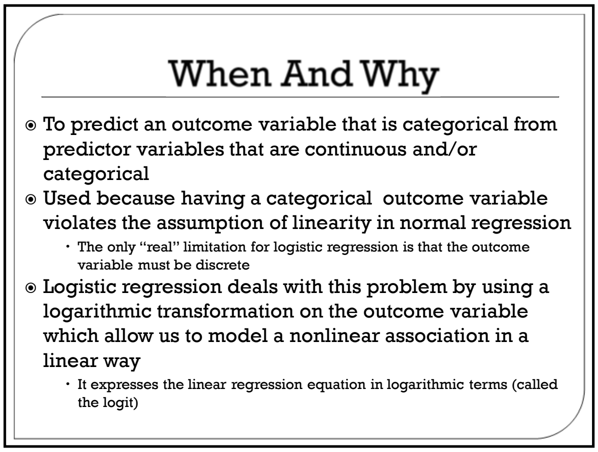
Logistic Regression

PSY 5102: Advanced Statistics for
Psychological and Behavioral Research 2



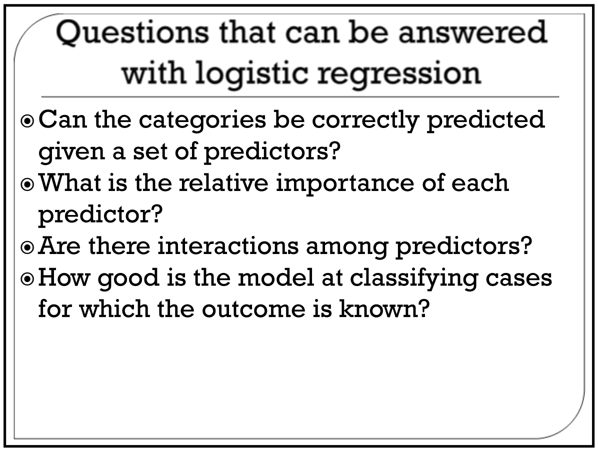
Goals

- When and why do we use logistic regression?
  - Binary
  - Multinomial
- Theory behind logistic regression
  - Assessing the model
  - Assessing predictors
  - Things that can go wrong
- Interpreting logistic regression



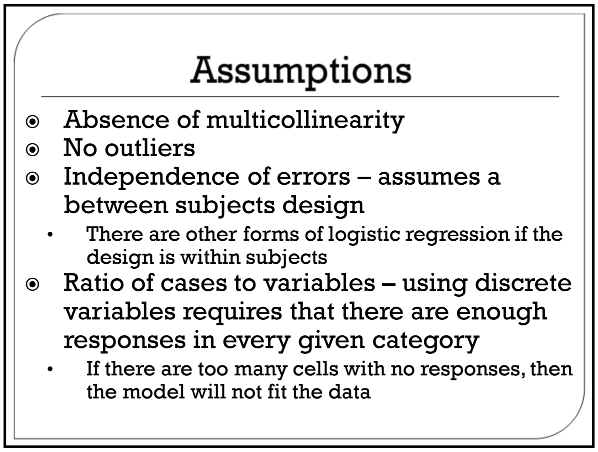
When And Why

- To predict an outcome variable that is categorical from predictor variables that are continuous and/or categorical
- Used because having a categorical outcome variable violates the assumption of linearity in normal regression
  - The only "real" limitation for logistic regression is that the outcome variable must be discrete
- Logistic regression deals with this problem by using a logarithmic transformation on the outcome variable which allow us to model a nonlinear association in a linear way
  - It expresses the linear regression equation in logarithmic terms (called the logit)



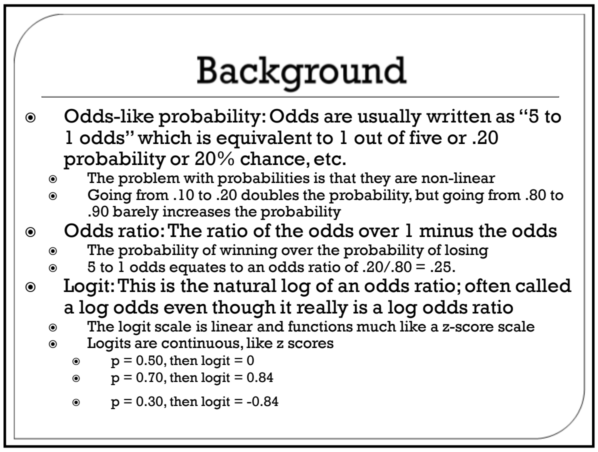
Questions that can be answered with logistic regression

- Can the categories be correctly predicted given a set of predictors?
- What is the relative importance of each predictor?
- Are there interactions among predictors?
- How good is the model at classifying cases for which the outcome is known?



Assumptions

- Absence of multicollinearity
- No outliers
- Independence of errors – assumes a between subjects design
  - There are other forms of logistic regression if the design is within subjects
- Ratio of cases to variables – using discrete variables requires that there are enough responses in every given category
  - If there are too many cells with no responses, then the model will not fit the data

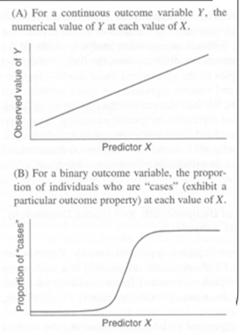


Background

- Odds-like probability: Odds are usually written as "5 to 1 odds" which is equivalent to 1 out of five or .20 probability or 20% chance, etc.
  - The problem with probabilities is that they are non-linear
  - Going from .10 to .20 doubles the probability, but going from .80 to .90 barely increases the probability
- Odds ratio: The ratio of the odds over 1 minus the odds
  - The probability of winning over the probability of losing
  - 5 to 1 odds equates to an odds ratio of .20/.80 = .25.
- Logit: This is the natural log of an odds ratio; often called a log odds even though it really is a log odds ratio
  - The logit scale is linear and functions much like a z-score scale
  - Logits are continuous, like z scores
    - p = 0.50, then logit = 0
    - p = 0.70, then logit = 0.84
    - p = 0.30, then logit = -0.84

Background

- An ogive function is a curved s-shaped function and the most common is the logistic function which looks like:



(A) For a continuous outcome variable *Y*, the numerical value of *Y* at each value of *X*.

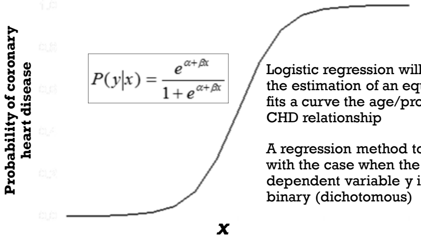
(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of *X*.

The Logistic Function

- $Y'_i = \frac{e^u}{1+e^u}$
- Where Y' is the estimated probability that the ith case is in a category and U is the regular linear regression equation:
- $U = A + B_1X_1 + B_2X_2 + \ldots + B_KX_K$

The Logistic Function

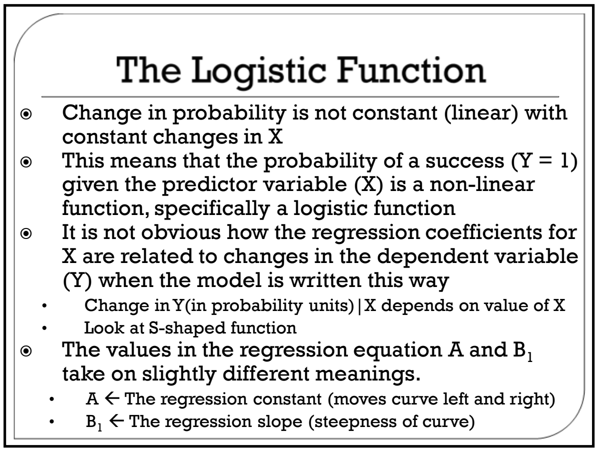
For a response variable y with p(y=1)= P and p(y=0) = 1- P



$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

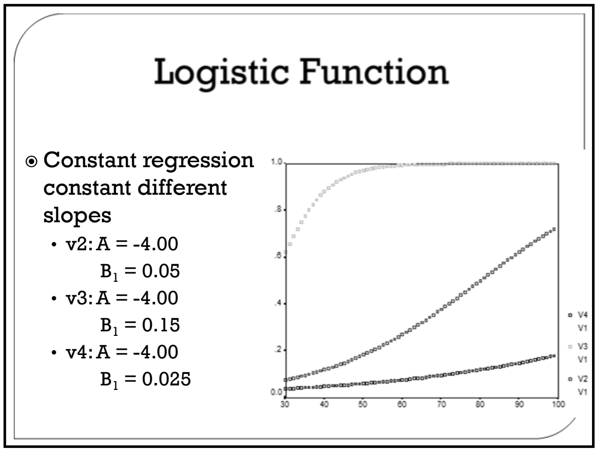
Logistic regression will allow for the estimation of an equation that fits a curve the age/probability of CHD relationship

A regression method to deal with the case when the dependent variable y is binary (dichotomous)



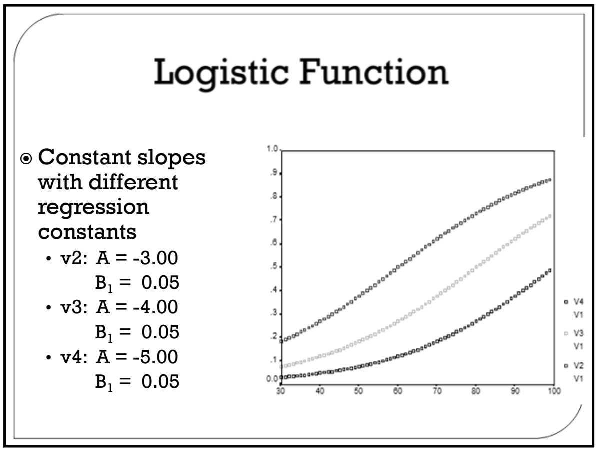
The Logistic Function

- Change in probability is not constant (linear) with constant changes in X
- This means that the probability of a success (Y = 1) given the predictor variable (X) is a non-linear function, specifically a logistic function
- It is not obvious how the regression coefficients for X are related to changes in the dependent variable (Y) when the model is written this way
  - Change in Y(in probability units)|X depends on value of X
  - Look at S-shaped function
- The values in the regression equation A and B_1 take on slightly different meanings.
  - A ← The regression constant (moves curve left and right)
  - B_1 ← The regression slope (steepness of curve)



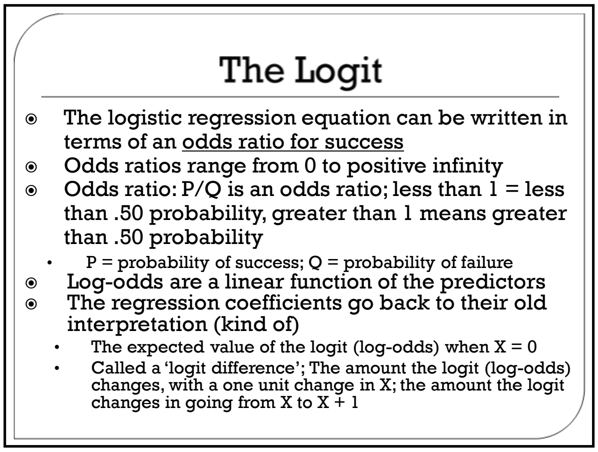
Logistic Function

- Constant regression constant different slopes
  - v2: A = -4.00
    B_1 = 0.05
  - v3: A = -4.00
    B_1 = 0.15
  - v4: A = -4.00
    B_1 = 0.025



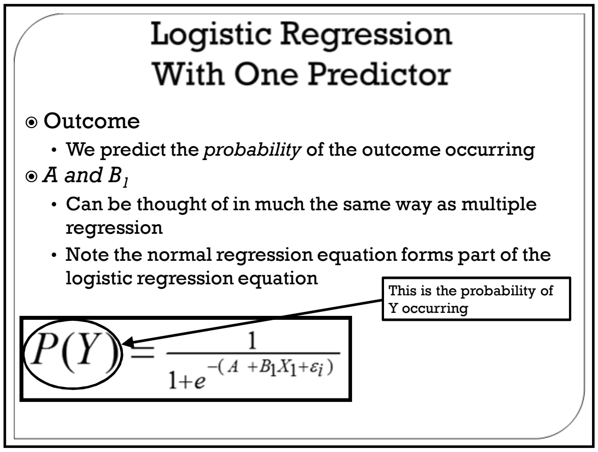
Logistic Function

- Constant slopes with different regression constants
  - v2: A = -3.00
    B_1 = 0.05
  - v3: A = -4.00
    B_1 = 0.05
  - v4: A = -5.00
    B_1 = 0.05



The Logit

- The logistic regression equation can be written in terms of an odds ratio for success
- Odds ratios range from 0 to positive infinity
- Odds ratio: P/Q is an odds ratio; less than 1 = less than .50 probability, greater than 1 means greater than .50 probability
  - P = probability of success; Q = probability of failure
- Log-odds are a linear function of the predictors
- The regression coefficients go back to their old interpretation (kind of)
  - The expected value of the logit (log-odds) when X = 0
  - Called a 'logit difference'; The amount the logit (log-odds) changes, with a one unit change in X; the amount the logit changes in going from X to X + 1

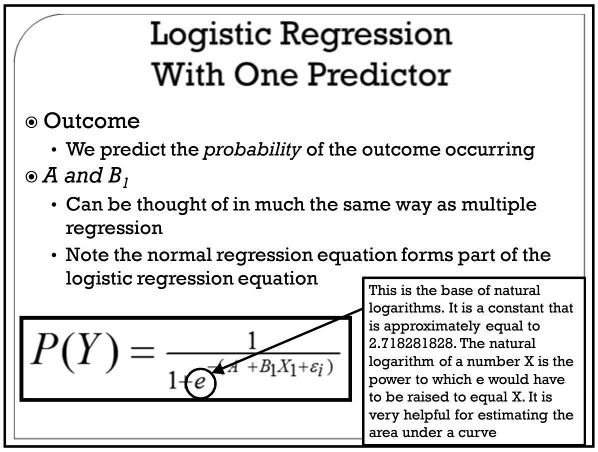


Logistic Regression With One Predictor

- Outcome
  - We predict the *probability* of the outcome occurring
- *A and B_1*
  - Can be thought of in much the same way as multiple regression
  - Note the normal regression equation forms part of the logistic regression equation

$$P(Y) = \frac{1}{1+e^{-(A + B_1X_1+\varepsilon_i)}}$$

This is the probability of Y occurring

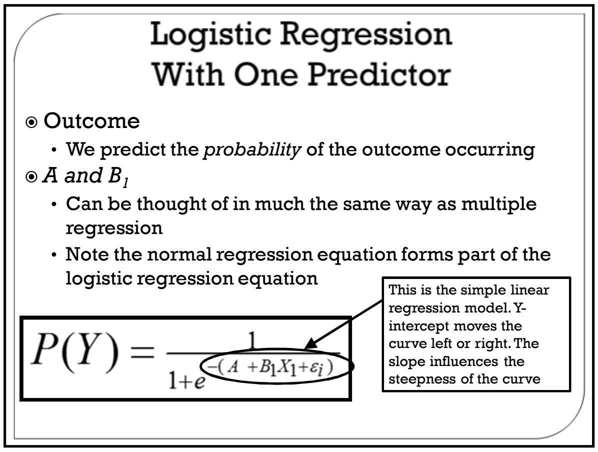


Logistic Regression With One Predictor

- Outcome
  - We predict the *probability* of the outcome occurring
- *A and B_1*
  - Can be thought of in much the same way as multiple regression
  - Note the normal regression equation forms part of the logistic regression equation

$$P(Y) = \frac{1}{1+e^{-(A + B_1X_1+\varepsilon_i)}}$$

This is the base of natural logarithms. It is a constant that is approximately equal to 2.718281828. The natural logarithm of a number X is the power to which e would have to be raised to equal X. It is very helpful for estimating the area under a curve

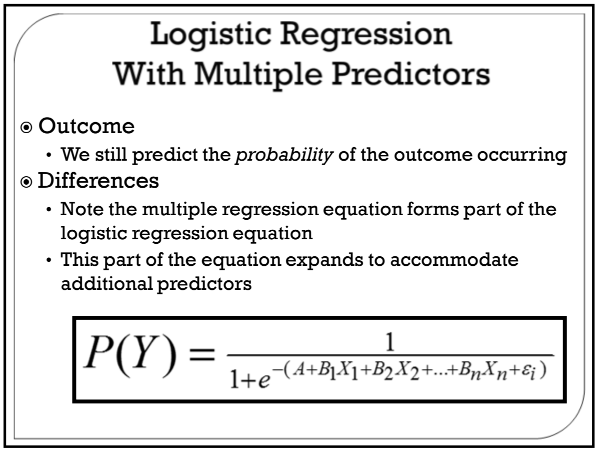


Logistic Regression With One Predictor

- ⦿ Outcome
  - We predict the *probability* of the outcome occurring
- ⦿ *A and B_1*
  - Can be thought of in much the same way as multiple regression
  - Note the normal regression equation forms part of the logistic regression equation

$$P(Y) = \frac{1}{1+e^{-(A\ +B_1X_1+\varepsilon_i)}}$$

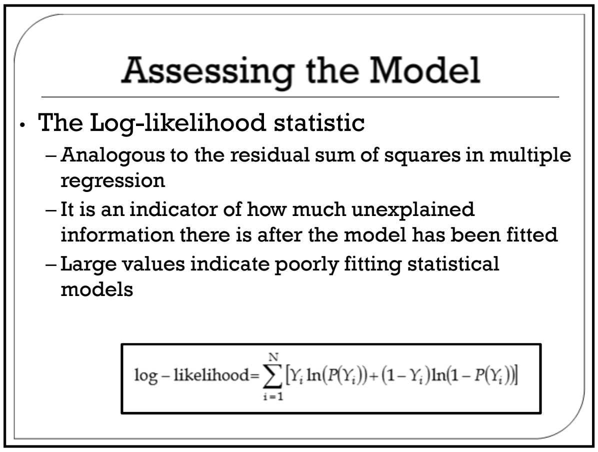
This is the simple linear regression model. Y-intercept moves the curve left or right. The slope influences the steepness of the curve



Logistic Regression With Multiple Predictors

- ⦿ Outcome
  - We still predict the *probability* of the outcome occurring
- ⦿ Differences
  - Note the multiple regression equation forms part of the logistic regression equation
  - This part of the equation expands to accommodate additional predictors

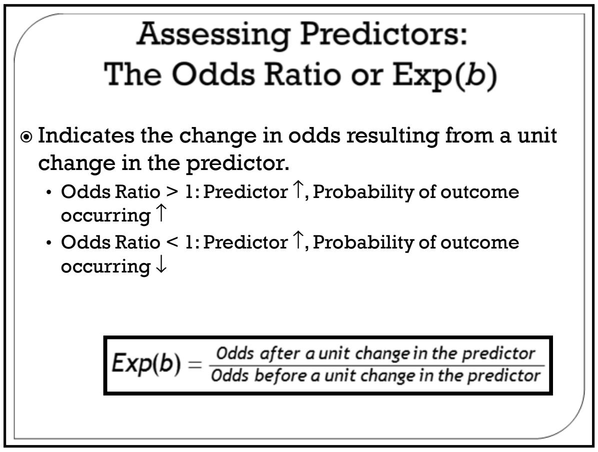
$$P(Y) = \frac{1}{1+e^{-(A+B_1X_1+B_2X_2+\ldots+B_nX_n+\varepsilon_i)}}$$



Assessing the Model

- The Log-likelihood statistic
  - Analogous to the residual sum of squares in multiple regression
  - It is an indicator of how much unexplained information there is after the model has been fitted
  - Large values indicate poorly fitting statistical models

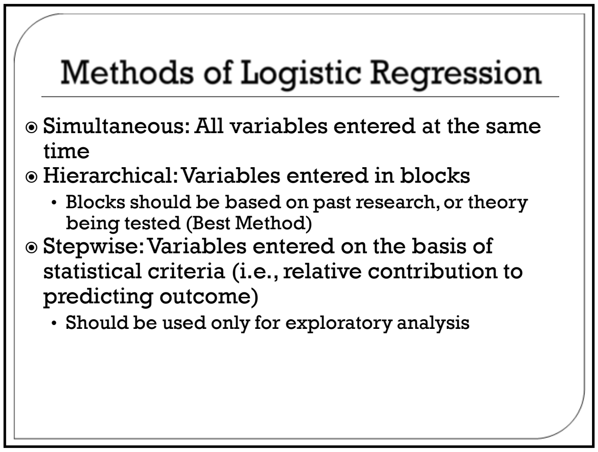
$$\log-\text{likelihood}=\sum_{i=1}^{N}\left[Y_i\,\mathrm{In}(P(Y_i))+(1-Y_i)\mathrm{In}(1-P(Y_i))\right]$$



Assessing Predictors: The Odds Ratio or Exp(*b*)

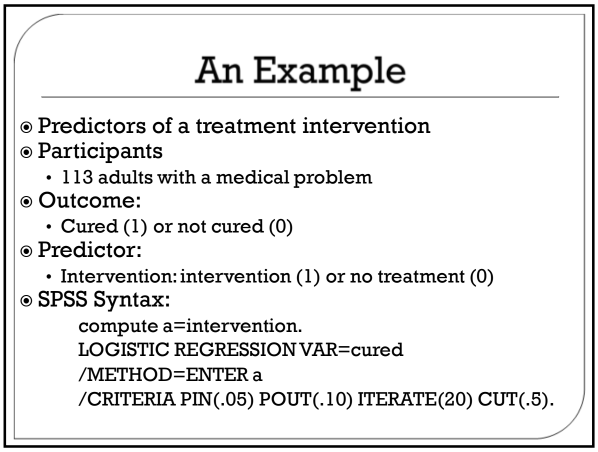
- Indicates the change in odds resulting from a unit change in the predictor.
  - Odds Ratio > 1: Predictor ↑, Probability of outcome occurring ↑
  - Odds Ratio < 1: Predictor ↑, Probability of outcome occurring ↓

$$Exp(b) = \frac{\text{Odds after a unit change in the predictor}}{\text{Odds before a unit change in the predictor}}$$



Methods of Logistic Regression

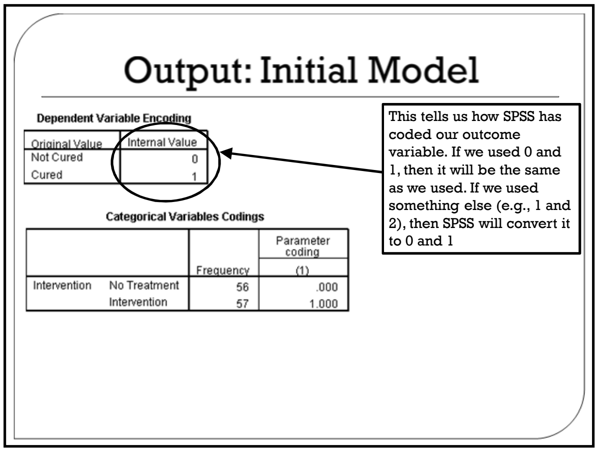
- Simultaneous: All variables entered at the same time
- Hierarchical: Variables entered in blocks
  - Blocks should be based on past research, or theory being tested (Best Method)
- Stepwise: Variables entered on the basis of statistical criteria (i.e., relative contribution to predicting outcome)
  - Should be used only for exploratory analysis



An Example

- Predictors of a treatment intervention
- Participants
  - 113 adults with a medical problem
- Outcome:
  - Cured (1) or not cured (0)
- Predictor:
  - Intervention: intervention (1) or no treatment (0)
- SPSS Syntax:

```
compute a=intervention.
LOGISTIC REGRESSION VAR=cured
/METHOD=ENTER a
/CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```



Output: Initial Model

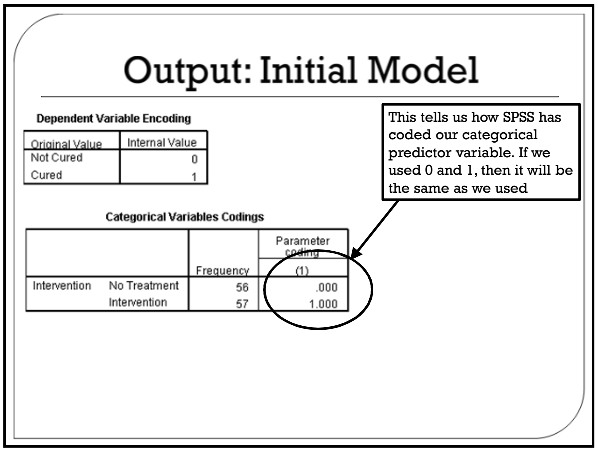
Dependent Variable Encoding

Original Value	Internal Value
Not Cured	0
Cured	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
Intervention	No Treatment	56	.000
	Intervention	57	1.000

This tells us how SPSS has coded our outcome variable. If we used 0 and 1, then it will be the same as we used. If we used something else (e.g., 1 and 2), then SPSS will convert it to 0 and 1



Output: Initial Model

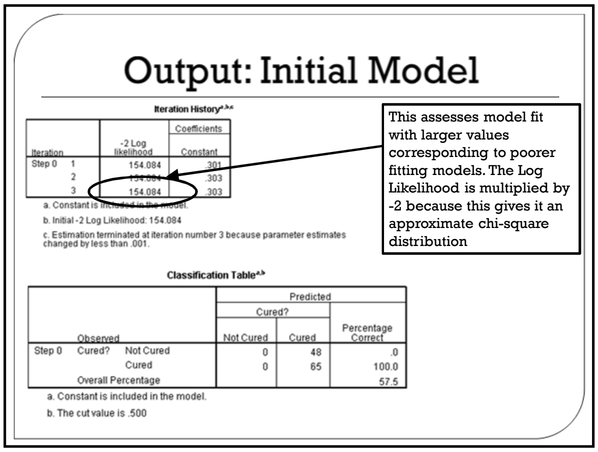
Dependent Variable Encoding

Original Value	Internal Value
Not Cured	0
Cured	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
Intervention	No Treatment	56	.000
	Intervention	57	1.000

This tells us how SPSS has coded our categorical predictor variable. If we used 0 and 1, then it will be the same as we used



Output: Initial Model

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	154.084	.301
	2	154.084	.303
	3	154.084	.303

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 154.084

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

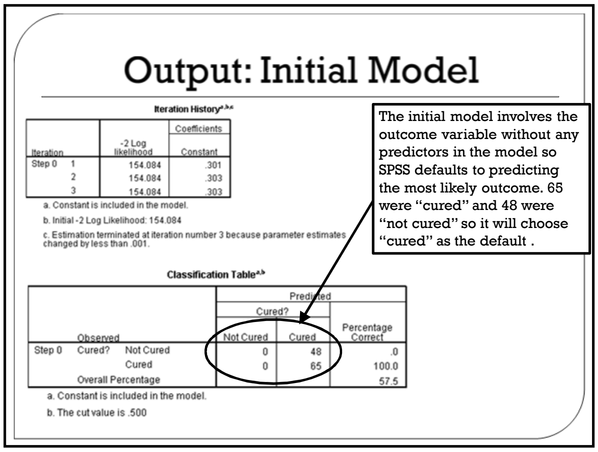
This assesses model fit with larger values corresponding to poorer fitting models. The Log Likelihood is multiplied by -2 because this gives it an approximate chi-square distribution

Classification Table[a,b]

Observed			Predicted Cured? Not Cured	 Cured	 Percentage Correct
Step 0	Cured?	Not Cured	0	48	.0
		Cured	0	65	100.0
	Overall Percentage				57.5

a. Constant is included in the model.

b. The cut value is .500



Output: Initial Model

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	154.084	.301
	2	154.084	.303
	3	154.084	.303

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 154.084

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

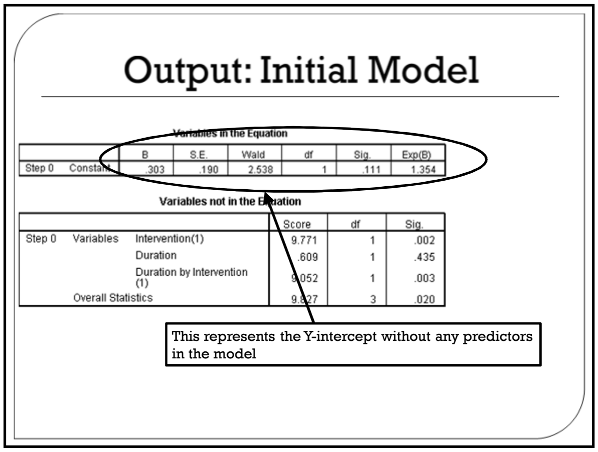
Classification Table[a,b]

Observed			Predicted		
			Cured?		Percentage Correct
			Not Cured	Cured	
Step 0	Cured?	Not Cured	0	48	.0
		Cured	0	65	100.0
	Overall Percentage				57.5

a. Constant is included in the model.

b. The cut value is .500

The initial model involves the outcome variable without any predictors in the model so SPSS defaults to predicting the most likely outcome. 65 were "cured" and 48 were "not cured" so it will choose "cured" as the default .



Output: Initial Model

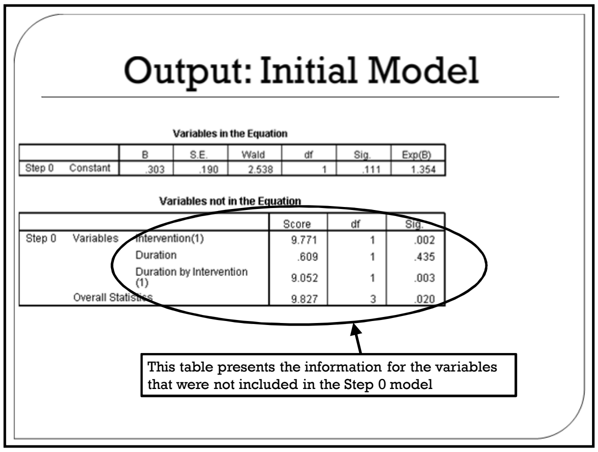
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.303	.190	2.538	1	.111	1.354

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Intervention(1)	9.771	1	.002
		Duration	.609	1	.435
		Duration by Intervention (1)	9.052	1	.003
	Overall Statistics		9.827	3	.020

This represents the Y-intercept without any predictors in the model



Output: Initial Model

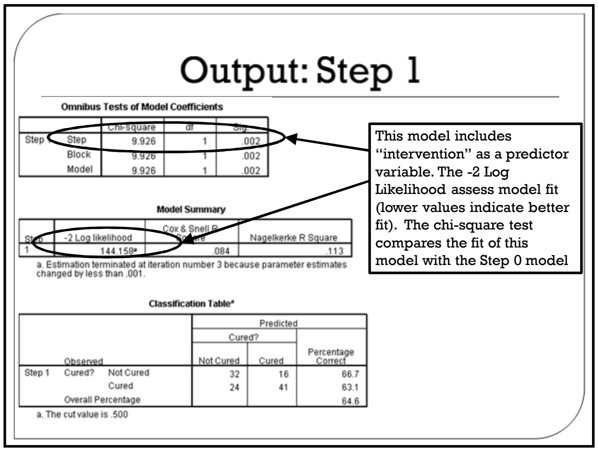
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.303	.190	2.538	1	.111	1.354

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Intervention(1)	9.771	1	.002
		Duration	.609	1	.435
		Duration by Intervention (1)	9.052	1	.003
	Overall Statistics		9.827	3	.020

This table presents the information for the variables that were not included in the Step 0 model



Output: Step 1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.926	1	.002
	Block	9.926	1	.002
	Model	9.926	1	.002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144.158[a]	.084	.113

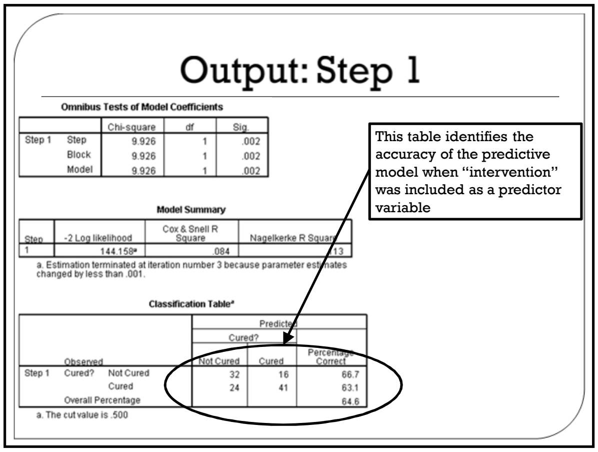
a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Cured?		Percentage Correct
Observed			Not Cured	Cured	
Step 1	Cured?	Not Cured	32	16	66.7
		Cured	24	41	63.1
	Overall Percentage				64.6

a. The cut value is .500

This model includes "intervention" as a predictor variable. The -2 Log Likelihood assess model fit (lower values indicate better fit). The chi-square test compares the fit of this model with the Step 0 model



Output: Step 1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.926	1	.002
	Block	9.926	1	.002
	Model	9.926	1	.002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144.158[a]	.084	.113

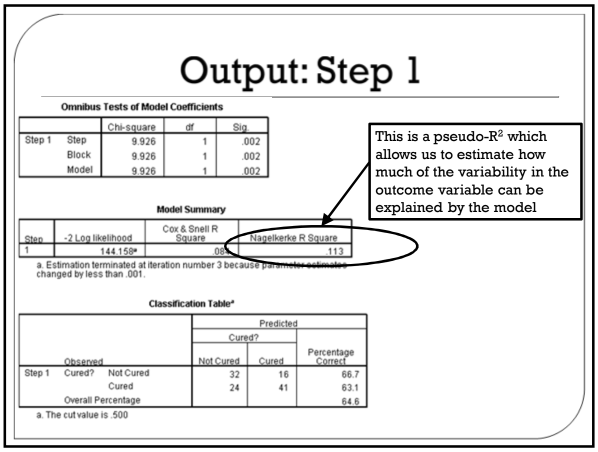
a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Cured?		Percentage Correct
Observed			Not Cured	Cured	
Step 1	Cured?	Not Cured	32	16	66.7
		Cured	24	41	63.1
	Overall Percentage				64.6

a. The cut value is .500

This table identifies the accuracy of the predictive model when "intervention" was included as a predictor variable



Output: Step 1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.926	1	.002
	Block	9.926	1	.002
	Model	9.926	1	.002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144.158[a]	.084	.113

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

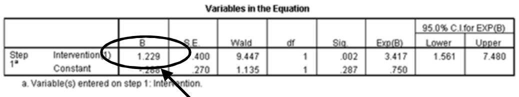
Classification Table[a]

			Predicted		
			Cured?		Percentage Correct
Observed			Not Cured	Cured	
Step 1	Cured?	Not Cured	32	16	66.7
		Cured	24	41	63.1
	Overall Percentage				64.6

a. The cut value is .500

This is a pseudo-R^2 which allows us to estimate how much of the variability in the outcome variable can be explained by the model

Output: Step 1



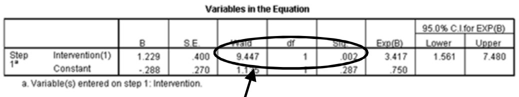
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1ª	Intervention(1)	1.229	.400	9.447	1	.002	3.417	1.561	7.480
	Constant	-.288	.270	1.135	1	.287	.750		

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

This value is the unstandardized regression coefficient that represents the slope of the model. It represents the change in the logit of the outcome variable (natural logarithm of the odds of Y occurring) associated with a one-unit change in the predictor variable

Output: Step 1



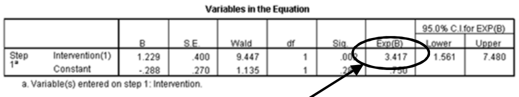
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1ª	Intervention(1)	1.229	.400	9.447	1	.002	3.417	1.561	7.480
	Constant	-.288	.270	1.135	1	.287	.750		

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

The Wald statistic is the crucial value because it tells us whether the B coefficient is significantly different from 0. If it is significantly different from 0, then we can assume that the predictor is making a significant contribution to the prediction of the outcome variable

Output: Step 1

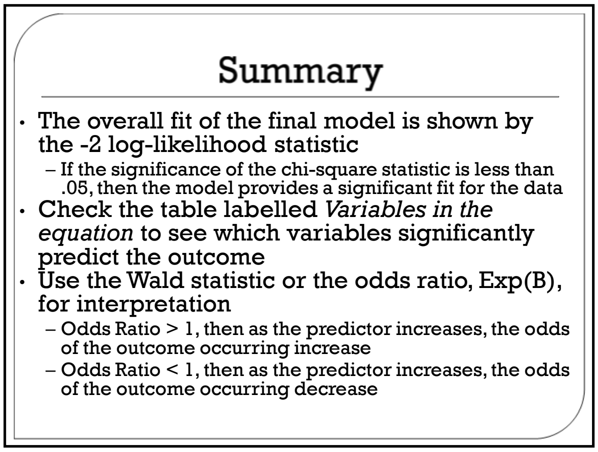


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1ª	Intervention(1)	1.229	.400	9.447	1	.002	3.417	1.561	7.480
	Constant	-.288	.270	1.135	1	.287	.750		

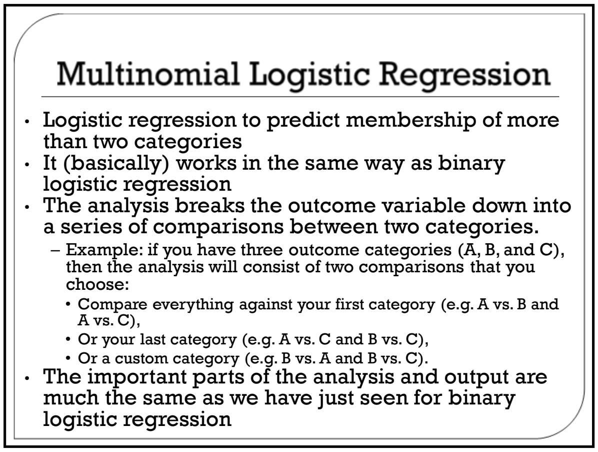
a. Variable(s) entered on step 1: Intervention.

This is the odds-ratio which is the odds (success) over 1 minus the odds (failure). In this example, we can say that the odds of a patient who is treated being cured are 3.41 times higher than those of a patient who is not treated



Summary

- The overall fit of the final model is shown by the -2 log-likelihood statistic
  - If the significance of the chi-square statistic is less than .05, then the model provides a significant fit for the data
- Check the table labelled *Variables in the equation* to see which variables significantly predict the outcome
- Use the Wald statistic or the odds ratio, Exp(B), for interpretation
  - Odds Ratio > 1, then as the predictor increases, the odds of the outcome occurring increase
  - Odds Ratio < 1, then as the predictor increases, the odds of the outcome occurring decrease



Multinomial Logistic Regression

- Logistic regression to predict membership of more than two categories
- It (basically) works in the same way as binary logistic regression
- The analysis breaks the outcome variable down into a series of comparisons between two categories.
  - Example: if you have three outcome categories (A, B, and C), then the analysis will consist of two comparisons that you choose:
    - Compare everything against your first category (e.g. A vs. B and A vs. C),
    - Or your last category (e.g. A vs. C and B vs. C),
    - Or a custom category (e.g. B vs. A and B vs. C).
- The important parts of the analysis and output are much the same as we have just seen for binary logistic regression