EXERCISE 3: SINGLE-SPECIES, SINGLE-SEASON OCCUPANCY MODELS (NO COVARIATES)

Please cite this work as: Donovan, T. M. and J. Hines. 2007. Exercises in occupancy modeling and estimation.
<http://www.uvm.edu/envnr/vtcfwru/spreadsheets/occupancy.htm>

TABLE OF CONTENTS

SINGLE-SPECIES, SINGLE-SEASON OCCUPANCY MODELS SPREADSHEET EXERCISE 3
OBJECTIVES: 3
BASIC INFORMATION 3
BACKGROUND 3
ENCOUNTER HISTORIES 4
MODEL ASSUMPTIONS 6
THE SATURATED MODEL 6
THE SATURATED MODEL'S MULTINOMIAL LOG LIKELIHOOD 7
OCCUPANCY MODEL PARAMETERS 9
LINKS 11
HISTORY PROBABILITIES AND MLE'S 14
MISSING SURVEYS 16
THE OCCUPANCY MODEL MULTINOMIAL LOG LIKELIHOOD 17
MAXIMIZING THE LOG LIKELIHOOD 18
MODEL RESULTS 20
ASSESSING FIT 21
COMPARING MODELS 25
SIMULATING DATA 27
SINGLE SEASON OCCUPANCY MODELS ANALYSIS IN PRESENCE 32
OBJECTIVES 32
DOWNLOADING PRESENCE 32
GETTING STARTED 32
THE PRESENCE INPUT FILE 33
MODEL PSI,P(T) 40
MODEL PSI,P(.) 47
OBJECTIVES 51
DOWNLOADING MARK 51
GETTING STARTED: CREATING AN INPUT FILE 51
MARK PIM (PARAMETER INDEX MATRIX) 54
MARK PIM CHART 56
THE RESULTS BROWSER IN MARK 60
THE FULL OUTPUT 61
THE PEARSON CHI-SQUARE OUTPUT 64
BETA AND REAL ESTIMATES 67
MODEL PSI(.)P(.) 69
MODEL PSI(.)P(.) OUTPUT 72

SINGLE-SPECIES, SINGLE-SEASON OCCUPANCY SPREADSHEET EXERCISE

OBJECTIVES:

- To learn and understand the basic occupancy model, and how it fits into a multinomial maximum likelihood analysis.
- To use Solver to find the maximum likelihood estimates for the probability of detection and the probability of site occupancy.
- To assess the $-2Log_eL$ of the saturated model.
- To introduce concepts of model fit.
- To learn how to simulate basic occupancy data.

BASIC INFORMATION

Now we turn to single season occupancy models, which follow the material by MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy when detection probabilities are less than one. Ecology 83:2248-2255. This article is on the web: http://www.uvm.edu/envnr/vtcfwru/spreadsheets/Occupancy/Occupancy3.htm and is also described in Chapter 4 of the book, "Occupancy Estimation and Modeling." Click on the worksheet labeled "Occupancy" and we'll get started.

BACKGROUND

For occupancy models, the goal is to determine the probability that a site is occupied, given that organisms are imperfectly detected. This is quite different than other mark-recapture analyses, in which the individual animal is the focus of attention. In this case, the focus is on a species of interest, the

occurrence of that species across different study sites (e.g. physical location), and the data are collected in a single-season.

Let's assume that you are interested in developing an occupancy model for a butterfly species. You select 250 study sites, and set out to survey each site three times in quick succession (or quick enough to assume that the site does not change in occupancy status between surveys). Or, alternatively, if the site is large enough, the site could have been sampled in three different locations in the same time period. In any event, the species is recorded as being present (1) or absent (0) from the site based on your survey efforts. Let's assume that on site 1 the butterfly was detected on the first survey, but not detected on the second or third survey. This is called an encounter history for site 1, and would be recorded as 100. In the same way, the encounter history for each site is determined. The single-season occupancy analysis focuses on the different kinds of encounter histories (which are like the plant phenotypes in the multinomial exercise) and their frequencies. In a nutshell, multinomial maximum likelihood procedures to estimate the key parameter of interest, psi (ψ), which is the probability that a site is occupied, as well as p_i (the probability of detecting the species on survey i, given the species is present on the site). We'll show you how this is done step by step.

ENCOUNTER HISTORIES

Let's review the notion of encounter histories first. We already mentioned that a history of 100 indicates the species was detected on the first survey, but not on the second or third survey. A history of 110 means that a species of interest was detected on the first and second samples, but not on sample 3. A

history of 010 means that the species was detected on the second sampling occasion, but not on the first or third occasion. A history of 000 means the species was not detected on any of the 3 sampling occasions. This does not mean the site was unoccupied; only that the species was not detected there. How many possible histories are there? Because a particular sample occasion has only two outcomes (0 or 1), and there are 3 sampling occasions, there are 2^3 = 8 kinds of histories in total. These histories written out in cells E4:E11. If there were four sampling occasions instead of three, there would be 2^4 = 16 possible encounter histories. If there were five sampling occasions, there would be 2^5 = 32 possible encounter histories. And if there were six sampling occasions, there would be 2^6 = 64 possible encounter histories.

Let's assume you sampled all 250 sites three times and recorded the frequency of each kind of capture history in cells F4:F11. For example, 22 sites had a history of '100', 73 sites had a history of '111', and so on. The total of all observations is computed in cell F12.

	E	F
3	History	Frequency
4	100	22
5	111	73
6	101	25
7	110	41
8	000	55
9	011	15
10	001	5
11	010	14
12	# Sites =	250
13	# Histories =	8

MODEL ASSUMPTIONS

Now that you understand the basics behind the data collection, it's time to review the assumptions of the single-season occupancy model and the conditions in which the data are collected. As explained in MacKenzie *et al.*, the assumptions of this model are as follows: 1) The system is demographically closed to changes in the occupancy status of site during the sampling period. At the species level, this means that a species cannot colonize/immigrate to a site, or go locally extinct/emigrate from that site during the course of the study. 2) Species are not falsely detected. 3) Detection at a site is independent of detection at other sites. This means that your sites should be far enough apart to be biologically independent. The multi-season occupancy model handles violations to demographic closure, and the misidentification model handles violations to false identification; both are discussed later in the book.

THE SATURATED MODEL

OK, now what? Well, as we've done with other exercises in the Spreadsheet Project, let's start by finding the probability of getting a 100, 111, 101, etc. history directly from the raw data. That is, we'll start with the saturated model. How do we compute these probabilities? Simple! Just divide the frequency of each history by the total number of sites, as done in cells N4:N11. For example, cell N4 has the formula =F4/$F$12, which is 22/250=0.088. Thus, the raw data tell us that the probability of getting a 100 history is 0.088. The probability of getting a 111 history is 73/250=0.292, and so on. These are simple proportions, and are the maximum likelihood probabilities based on raw

data. You can't get better probability estimates than this….they "fit" the data perfectly.

	E	F	N	O
2			**Saturated Model**	
3	History	Frequency	Probability	Ln(Prob)
4	100	22	0.088	-2.43041846
5	111	73	0.292	-1.23100148
6	101	25	0.1	-2.30258509
7	110	41	0.164	-1.80788885
8	000	55	0.22	-1.51412773
9	011	15	0.06	-2.81341072
10	001	5	0.02	-3.91202301
11	010	14	0.056	-2.88240359
12	# Sites =	250	Log L (sat)	-460.41234
13	# Histories =	8	-2 Log L (sat)	920.82467

THE SATURATED MODEL'S MULTINOMIAL LOG LIKELIHOOD

OK, now what? Well, remember that basis for almost all the analyses in MARK or PRESENCE is the multinomial maximum likelihood formula. We assume that you've already completed the maximum likelihood spreadsheet exercise, but if you're rusty, take time to review it now. The multinomial likelihood function has the form:

$$L(\boldsymbol{p}_i \mid \boldsymbol{n}_i, \boldsymbol{y}_i) = \binom{\boldsymbol{n}}{\boldsymbol{y}_i} \boldsymbol{p}_1^{y1} \boldsymbol{p}_2^{y2} \boldsymbol{p}_3^{y3} \ldots \boldsymbol{p}_8^{y8}$$

This function states that the likelihood of p_i (the probability of encounter history i), given n (the total number of sites sampled) and y_i (the frequency of each type of encounter history) is equal to the multinomial coefficient $\binom{n}{y_i}$ multiplied by the probability of history 1 raised to its frequency, multiplied

by the probability of history 2, raised to its frequency, etc. etc. until all 8 histories are accounted for. (Remember, our spreadsheet example has 8 different histories). Commonly, the natural log version of this formula is used and the multinomial coefficient is dropped because it doesn't contain any parameters of interest. The log likelihood function has the form:

$$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + ... + y_8 \ln(p_8)$$

This function says that the log likelihood of the parameters, p_i, given n (the total sites sampled) and y_i (the frequency of each history) is <u>proportional</u> to y_1 times $\ln(p_1)$ + y_2 times $\ln(p_2)$ + …y_8 times $\ln(p_8)$.

So, what is the log likelihood of the saturated model? Well, we know the frequencies for each history (cells F4:F11), and we just computed the probabilities for each history based on the raw data (cells N4:N11). Now we need to take the natural logs (ln) of each probability (as done in cells O4:O11), and then multiply each history by the natural log of its probability, and add it all up. This is done in cell O12 with the formula =SUMPRODUCT(F4:F11,O4:O11), which is a pretty handy Excel function. This equation gives the same result as the long-hand function =F4*O4+F5*O5+F6*O6+F7*O7+F8*O8+F9*O9+F10*O10+F11*O11. The result is the saturated model's Log_eL. If we multiply cell O12 by -2, the result is the saturated model's $-2Log_eL$ (as shown in cell O13). What's the big deal? Well, because the probabilities for each history were computed from the raw data, the saturated model fits the data perfectly. The saturated model's $-2Log_eL$ therefore sets the "standard" upon which other models will be based.

	E	F	N	O
2			**Saturated Model**	
3	History	Frequency	Probability	Ln(Prob)
4	100	22	0.088	-2.43041846
5	111	73	0.292	-1.23100148
6	101	25	0.1	-2.30258509
7	110	41	0.164	-1.80788885
8	000	55	0.22	-1.51412773
9	011	15	0.06	-2.81341072
10	001	5	0.02	-3.91202301
11	010	14	0.056	-2.88240359
12	# Sites =	250	Log L (sat)	-460.41234
13	# Histories =	8	-2 Log L (sat)	920.82467

OCCUPANCY MODEL PARAMETERS

Now that we know the saturated model's $-2\text{Log}_e\text{L}$, we can forge ahead and estimate the probability of getting a 111, 110, etc., etc, not by the raw data, but by estimating the occupancy model parameters: p_1, p_2, p_3 and ψ. What are these? Since we surveyed our sites 3 times, there are three detection parameters (p_1, p_2, and p_3), and psi (or ψ), which is the actual probability that the site is occupied. (It's a bit unfortunate that the detection parameters are labeled p_1, p_2, and p_3 because it's easy to confuse these with the p_i's in the multinomial function. Just keep in mind that when we say p_1, p_2, or p_3 we are usually referring to the probability of detecting a species during occasion 1, 2, or 3, unless otherwise noted.) In a nutshell, we will estimate the probability of getting a particular history by combining estimates of p_1, p_2, p_3 and ψ, rather than estimating the probabilities from the raw data as we did for the saturated model. For instance, the probability of observing a 111 history is $\psi * p_1 * p_2 * p_3$ – we know the site was occupied because the species was observed on at least one occasion (ψ), and it was detected on the first (p_1), second (p_2), and third survey

(p_3). We'll cover the encounter history probabilities in a few minutes…for now our focus is on how to estimate ψ, p_1, p_2, and p_3. We don't need to worry about overparameterizing this model because only 4 parameters are being estimated (at the most), whereas the saturated model estimates 8 parameters (see Burnham and Anderson for a discussion of overparameterization if this may be a problem).

OK, so let's define the occupancy parameters more formally: p_1 is the probability that an animal is detected at a site during session 1, given that the site is actually occupied; p_2 is the probability that an animal is detected at a site during session 2, given that the site is actually occupied; p_3 is the probability that an animal is detected at an occupied site during occasion 3, and psi (ψ) is the probability that a site is occupied by the species of interest. Two things you should keep in mind. First, probabilities are bounded between 0 and 1. Second, remember that our population is assumed to be closed, so that if the species was detected in one survey, the species must have also been present in other surveys, even if it was not detected.

	G	H	I	J
3	Parameter	Estimate?	Betas	MLE
4	p1	1		0.50000
5	p2	1		0.50000
6	p3	1		0.50000
7	ψ	1		0.50000

Now let's return to the spreadsheet. In this spreadsheet, the occupancy parameters are listed in cells G4:G7. If you are going to estimate a certain parameter, enter a 1 in its corresponding column to the right (cells H4:H7). And if you're going to force a parameter to be equal to some other parameter, enter a 0 in its corresponding column to the right. (This will help you keep track of

the number of parameters you are estimating for any given model run….we'll get to this in a bit, so don't worry if this is confusing right now.) The set-up shown above, where H4:H7 = 1, indicates that we are interested in running a model where we separately estimate p_1, p_2, p_3, and ψ.

Our goal is to estimate p_1, p_2, p_3, and ψ, so that we can then use the estimates to derive the probability of observing each of the 8 different encounter histories within the multinomial function. But, here's a little catch: we won't estimate p_1, p_2, p_3, and ψ directly – instead we'll estimate their corresponding beta values. The betas are values in cells I4:I7. We don't know what they are…we'll let Solver find them for us. Why use the betas? Well, probabilities are constrained between 0 and 1, whereas the betas are not. The betas are used because it's easier for Solver to find betas that are unconstrained, and also they are needed for more advanced modeling in which we force a parameter to be a function of covariates (which we'll cover in the next exercise). After finding a particular beta value, we use a "link function" to convert a beta value for a parameter back to its corresponding probability.

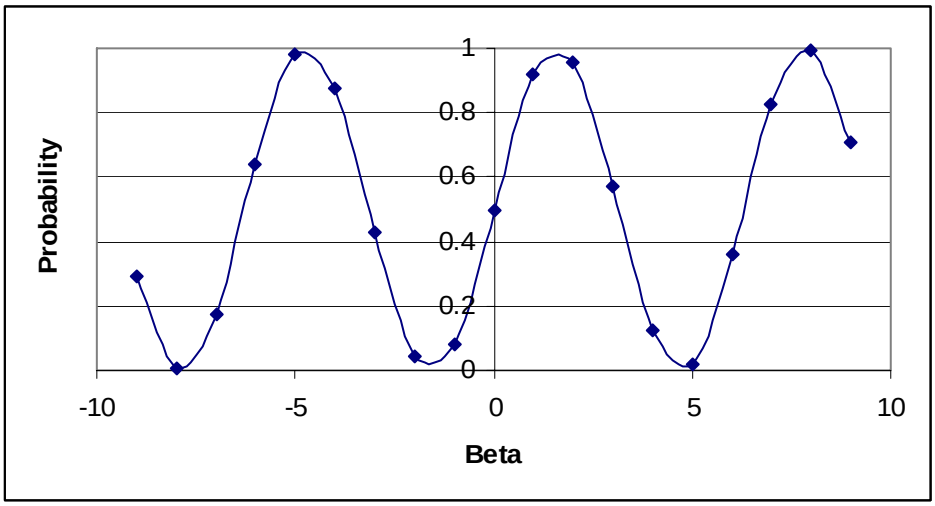
LINKS

The betas and actual MLE's are linked together with a sin link, as you can see by looking at the formulae in cells J4:J7. For example, enter the formula =(SIN(I4)+1)/2 in cell J4, and copy it down to cell J7. A "link" function simply converts the betas (which can be any number) to a probability (which is bound between 0 and 1). There are many different kinds of link functions; we'll start to the sin link in this exercise, but will quickly replace it with the logit link. The sin link function has the form: probability = (sin(beta)+1)/2. In cell J4, enter

the formula =(sin(I4)+1)/2, and copy this formula down to cell J7. Note that the betas can take on any value, and the link function constrains the MLE's to be between 0 and 1, which is necessary because p_1, p_2, p_3, and ψ are probabilities, and probabilities range between 0 and 1. In the example below, we entered beta values of -1, 0, 1, and 2 for p_1, p_2, p_3, and ψ,

	G	H	I	J
3	Parameter	Estimate?	Betas	MLE
4	p1	1	-1	0.07926
5	p2	1	0	0.50000
6	p3	1	1	0.92074
7	ψ	1	2	0.95465

respectively, and these correspond to the following probabilities: p_1 = 0.07926, p_2 = 0.5000, p_3 = 0.92074, and ψ = 0.95465. The figure below shows betas that range from -9 to +9, and the corresponding, sin "transformed" probability estimate. Look for the beta values of -1, 0, 1, and 2, and find their corresponding probabilities on the graph:

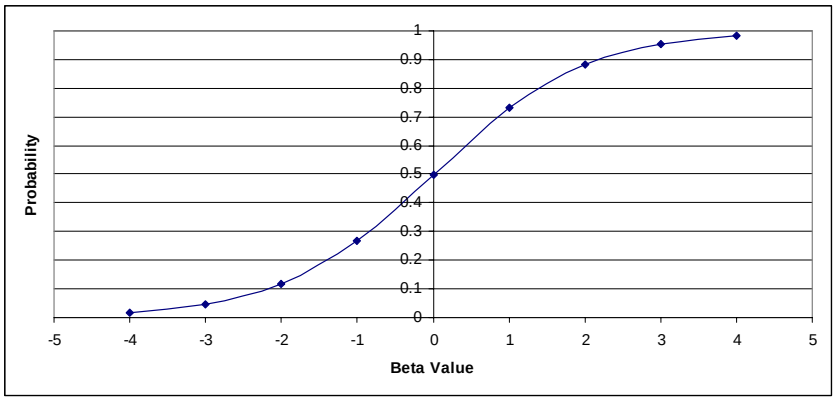


Now enter some numbers of your choice in the beta column (cells I4:I7) and examine the MLE's. You should see that no matter what beta values you enter, the corresponding parameters are constrained between 0 and 1.

The other link of interest is the logit link, which we'll use almost exclusively for the remainder of the exercises in this book. In cell J4, enter the equation =EXP(I4)/(1+EXP(I4)), and then copy this formula down for the other parameters (cells J5:J7). Try it; you should get the following parameter estimates for the betas entered:

	G	H	I	J
3	Parameter	Estimate?	Betas	MLE
4	p1	1	-1	0.26894
5	p2	1	0	0.50000
6	p3	1	1	0.73106
7	ψ	1	2	0.88080

Notice that different links provide different estimates for the same beta value, and some links are more appropriate than others depending on your analysis. A graph of beta values and their corresponding logit-link probabilities is shown below. This logit link function is said to be monotonically increasing – for every beta value there is a single probability. This graph highlights the fact that betas less than -4 correspond to a probability near 0, and betas greater than +4 correspond to a probability near 1. In contrast, several different beta values correspond to the same probability value when the sin link function is used.



HISTORY PROBABILITIES AND MLE'S

So, given a set of betas, we derive estimates for p_1, p_2, p_3, and ψ. The probability cells are "named". For example, cell J4 is named _p1, cell J5 is named _p2, cell J6 is named _p3, and cell J7 is named psi. (Click on a cell and notice the name of that cell appears to the left of the formula bar).

Microsoft Excel - Occupancy Spreadsheet Exercises.xls

File Edit View Insert Format Tools PopTools Data Window Help Adobe PDF

100% Comic Sans MS

_p1 =EXP(I4)/(1+EXP(I4))

	A	B	C	D	E	F	G	H	I	J	K
1					Single-Species, Single-Season Mode						
2											Probability
3					History	Frequency	Parameter	Estimate?	Betas	MLE	of History
4					100	22	p1	1	-1	0.26894	0.032
5					111	73	p2	1	0	0.50000	0.087
6					101	25	p3	1	1	0.73106	0.087
7					110	41	ψ	1	2	0.88080	0.032
8					000	55					0.206
9					011	15					0.235
10					001	5					0.235
11					010	14					0.087
12					# Sites =	250					1
13					# Histories =	8					

It would have been more sensible to name them p1, p2, and p3, but those names are invalid in Excel because they refer to the location of a cell on the spreadsheet.

As you can see from the diagram above, the next step is to compute the probability of realizing each encounter history, given the betas and their corresponding parameter estimates. Let's start with the 100 history and use the occupancy model parameters to write an equation for this history's probability in cell K4. In this case, the species was detected on the first sampling occasion only. You can see that, given the betas entered, the probability of getting a 100 history is 0.032. How did we get that? Well, because it was detected at least once, we KNOW that the site was occupied, but it was detected on only one of the three sampling occasions. So the probability of its history is the probability that it was actually on the site (ψ) times the probability that it was detected on the first sampling occasion (_p1), times the probability it was missed on the second occasion (1-_p2), times the probability it was missed on the third occasion (1-_p3). This is the equation in cell K4 =psi*_p1*(1-_p2)*(1-_p3), which generates the probability of getting a 100 history. Given the betas entered, the probability of realizing a 100 history is 0.032. The next history is 111. This is computed as =psi*_p1*_p2*_p3, which, given the betas entered, is 0.087. Pretty straight-forward. Let's try one more, history 101. We KNOW the site was occupied because the species was detected in at least one survey, and we know it was detected on the first and third survey but not the second. The probability of getting a 101 history is therefore =psi*_p1*(1-_p2)*_p3 (cell K6). Go ahead and work your way through the remaining histories. Don't skip this step! There's something "healthy" about writing out encounter history probabilities - it forces the concepts of the model to really sink in. The only tricky history is 000.

The 000 history is "tricky" only in that there are 2 ways in which this history could be generated: the species could have been present on the site but missed in all three sampling occasions, which is $\psi*(1-p_1)*(1-p_2)*(1-p_3)$, OR the site really could have been unoccupied, which is $(1-\psi)$. So the probability of getting a 000 history needs to account for both possibilities, and so we add the two alternatives: =psi*(1-_p1)*(1-_p2)*(1-_p3) + (1-psi). With the betas entered as shown above, the probability is 0.206. Got it? The sum of the capture histories must equal 1, which is double-checked in cell K12.

Note: If you are a programmer, you wouldn't enter the history probabilities as we just did…things can get out of hand very quickly as the number of sites increases. Instead, you would use matrices to derive the history probabilities. We'll use that kind of approach when we pursue Study Designs.

MISSING SURVEYS

As a side note, it's often the case in field research where, for some reason or another, you are not able to survey a site. It's one thing to design a study on paper; it's another thing to carry the plans out! If you miss a survey at a site, you can still use the data from that site. For example, let's assume you were able to survey on the first and third occasion, but missed the second occasion as planned. If you detected the species on the first occasion, skipped the second survey, and failed to detect the species on the third occasion, the encounter history would be 1.0, where the dot indicates a missed survey. The probability of this history would be $\psi*p_1*(1-p_3)$. You certainly should keep the site in the analysis because this site provides information on ψ, p_1, and p_3, even

though it does not provide any information on p_2. More information about handling missing data is provided later in the book.

THE OCCUPANCY MODEL MULTINOMIAL LOG LIKELIHOOD

OK, where are we headed? Given the betas you entered, and hence the probabilities for p_1, p_2, p_3, and ψ (labeled as MLE's even though they are not yet maximized), and hence the probabilities of the various histories, we can compute the multinomial likelihood of observing the data we observed (cell E17). For the betas we entered, the model's Log_eL is –607.04. Don't worry about columns L through O for now…we'll get there soon enough.

	E	F	G	H	I	J	K	L	M	N	O
1	**Single-Species, Single-Season Model**										
2							Probability			**Saturated Model**	
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History	Expected	Chi-Square	Probability	Ln(Prob)
4	100	22	p1	1	-1	0.26894	0.032	7.96	24.74	0.088	-2.43041846
5	111	73	p2	1	0	0.50000	0.087	21.65	121.83	0.292	-1.23100148
6	101	25	p3	1	1	0.73106	0.087	21.65	0.52	0.1	-2.30258509
7	110	41	ψ	1	2	0.88080	0.032	7.96	137.05	0.164	-1.80788885
8	000	55					0.206	51.45	0.25	0.22	-1.51412773
9	011	15					0.235	58.84	32.67	0.06	-2.81341072
10	001	5					0.235	58.84	49.27	0.02	-3.91202301
11	010	14					0.087	21.65	2.70	0.056	-2.88240359
12	# Sites =	250					1	250	369.0186943	Log L (sat)	-460.41234
13	# Histories =	8								-2 Log L (sat)	920.82467
14											
15	**OUTPUTS**										
16	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-607.04	1214.072053	4	1222.07	1222.24	920.8247	293.2474	4	73.31184543	369.0187	0.0000

Now, in case you haven't guessed already, a major goal of the analysis is to find the combination of betas (and corresponding MLE's for p_1, p_2, p_3, and ψ) that maximize the Log_eL. In doing so, we find encounter history probabilities that will most closely match the observed frequencies in cells F4:F11. So the order of operation is really:

Betas → MLEs → Encounter History Probabilities → Multinomial Likelihood.

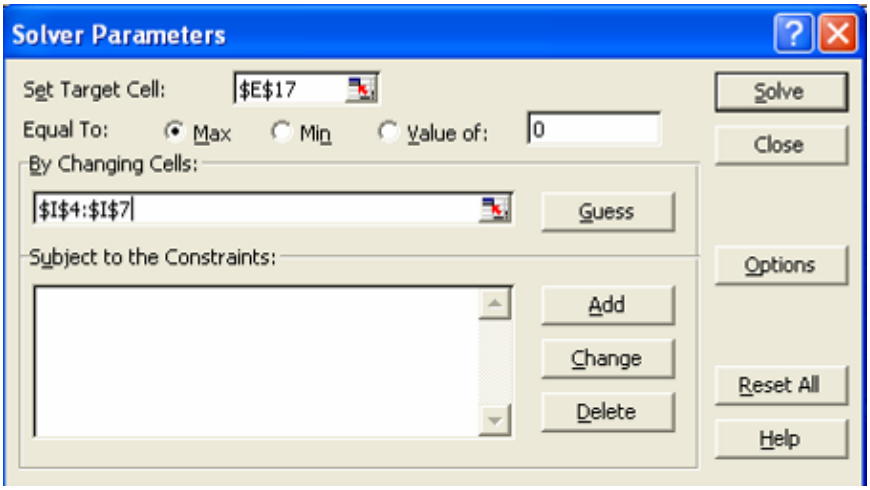
In other words, we need to find the betas, and hence the MLE's, which influence the encounter history probabilities, which maximize the likelihood of observing the data. The multinomial log likelihood is computed in cell E17, and is based on the frequencies of each kind of history and their corresponding capture history probability.

Click on cell E17, then place your mouse on the formula bar and click again……you should see the cells used in the formula "light up", revealing the pattern of the equation. This should look familiar to you by now:

$$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + ... + y_8 \ln(p_8)$$

MAXIMIZING THE LOG LIKELIHOOD

The goal now is to maximize the log likelihood (cell E17) by changing the betas (which are linked to the MLE's), and you can use the Solver to do this for you. (If Solver is not installed on your spreadsheet, go to Tools | Add-Ins, and select the Solver add-in.) In this example, we are interested in obtaining estimates for p_1, p_2, p_3, and ψ, so we enter a 1 in cells H4:H7. The total number of parameters you are trying to estimate (K) sums these 1's, and is computed in cell G17 with a SUM function. Now, run Solver. Go to Tools | Solver, and a dialogue box should appear. You want to set cell E17 to a maximum by changing cells I4:I7. You can either type in the cell references, or click on the small red arrow to the right of the box (which shrinks the dialogue box and reveals your spreadsheet), and then highlight the cells with your mouse. In this case there are no constraints.



The press Solve, and the Solver will search for the beta combinations, and hence combinations of p_1, p_2, p_3, and ψ, that maximize the multinomial log likelihood. When Solver finds its solution, click OK to keep the Solver results. If you run this analysis in MARK or PRESENCE, the MLE's and betas should match, provided you use the logit link. In MARK language, you'd call this model psi(.), p(t), indicating that you want separate estimates for psi and for <u>each</u> of the three p's. That's really all there is to it (in terms of the essential model basics). Here is the output we got (below).

	E	F	G	H	I	J	K
2							Probability
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History
4	100	22	p1	1	1.432238014	0.80725	0.074
5	111	73	p2	1	0.929620335	0.71700	0.273
6	101	25	p3	1	0.37078615	0.59165	0.108
7	110	41	ψ	1	1.372416073	0.79777	0.189
8	000	55					0.220
9	011	15					0.065
10	001	5					0.026
11	010	14					0.045
12	# Sites =	250					1
13	# Histories =	8					

MODEL RESULTS

Let's take a closer look at the model results. Solver changed the beta in cells I4:I7 to yield the following MLE's in J4:J7: $p_1 = 0.807$, $p_2 = 0.717$, $p_3 = 0.592$, and $\psi = 0.798$. If we had used a sin link instead of the logit link, the beta values would be different, but the MLE's would have been identical. So, according to this particular model, our butterfly species had a decreasing detection probability from survey 1 to survey 3. The probability of site occupancy was 0.798, indicating that almost 80% of the sites were occupied by the butterfly species of interest. You'll see later in the exercise that the data were simulated such that $p_1 = 0.8$, $p_2 = 0.7$, $p_3 = 0.6$, and $\psi = 0.8$, so Solver found estimates that were <u>unbiased</u>. Notice, however, that we can't tell how <u>precise</u> these estimates are….the estimates of standard errors around each parameter have not been added to the spreadsheet yet, but it's important that you examine the value of the estimate itself AND its precision when you interpret results.

Now let's look at the remaining model output:

	E	F	G	H	I	J	K	L	M	N	O
15	OUTPUTS										
16	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-461.89	923.7841681	4	931.78	931.95	920.8247	2.9595	4	0.73987425	2.9792	0.5613

The model's Log_eL is given in cell E17. Remember, this is the cell we maximized. You could change the beta values in cells I4:I7 to your heart's content, but you won't find a combination that will give you a lower value Log_eL for this model than -461.89 (except in cases where Solver potentially finds a <u>local</u> maximum…..see

http://www.uvm.edu/envnr/vtcfwru/spreadsheets.html#Maximum%20Likelihood%20Procedures for more details. This won't be a concern for the datasets we'll be analyzing in this exercise). The $-2Log_eL$ is given in cell F17, which is simply the Log_eL multiplied by -2. K is the number of parameters estimated in this model, and is 4 because we estimated p_1, p_2, p_3, and ψ separately -- the equation is in cell G17 is =SUM(H4:H7). AIC is the $-2Log_eL + 2K$, and is computed in cell H17, and AICc (AIC corrected for sample size, which in this case is the number of sites) is computed in cell I17. The AIC and AICc scores mean little unless you use them to compare a variety of models against each other…we'll come back to that in a minute. The model's degree of freedom (or deviance degrees of freedom) is given in cell L17, which is simply the number of histories minus K. The concept of model degrees of freedom refers to the difference in the number of parameters of the saturated model and the number of parameters in the occupancy model of interest.

ASSESSING FIT

OK, but just how valid are these model results? Good question! There are several ways to assess "fit", and we'll look at a couple of different approaches right now. First, this model's deviance is measured as the difference between the saturated model's $-2Log_eL$ and this model's $-2Log_eL$. Remember how we computed the $-2Log_eL$ for the saturated model (cell O13)? Well, deviance is computed in cell K17 as =F17-O13. Remember that the saturated model fits the data perfectly (by design), so this particular model "deviates" from the saturated model by 2.9595 units. A small number here is better because the closer a model is to the saturated model, the better it explains the field data.

How to interpret deviance isn't nearly as intuitive as other methods, so let's now look at the Pearson Chi-Square. As a quick summary, we ran a model to find those combinations of betas, and hence MLE's, which maximized the multinomial log likelihood. The probability of getting each history is obtained by plugging the MLE's into the history probability equation. For example, the probability of getting a 100 history is psi*p_1*(1-p_2)*(1-p_3) = 0.7977*0.80725*(1-0.71700)*(1-0.59165) = 0.074, which is given in cell K4. If this is the probability of getting a 100, and if we surveyed 250 sites, then we would expect that 250*0.074 = 18.61 sites to have a 100 history. This is exactly what was done in cell L4, which has the equation =$F$12*K4.

	E	F	G	H	I	J	K	L	M
2							Probability		
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History	Expected	Chi-Square
4	100	22	p1	1	0.661751265	0.80725	0.074	18.61	0.62
5	111	73	p2	1	0.448923955	0.71700	0.273	68.30	0.32
6	101	25	p3	1	0.184340048	0.59165	0.108	26.96	0.14
7	110	41			2.503652838	0.79777	0.189	47.14	0.80
8	000	55					0.220	55.00	0.00
9	011	15					0.065	16.31	0.10
10	001	5					0.026	6.44	0.32
11	010	14					0.045	11.26	0.67
12	# Sites =	250					1	250	2.979236969
13	# Histories =	8							

Similarly, given the MLE's found by Solver, the probability of getting a 111 history is ψ*p_1*p_2*p_3 = 0.7977*0.80725*0.71700*0.59165 = 0.273, which is computed in cell K5. Given this probability, we would expect 250 * 0.273 = 68.30 sites to have this kind of history (cell L5). Cells L4:L11 are therefore the number of sites expected to demonstrate each history, given the model's MLE's and the total sites sampled. (By the way, this is a handy way to make up datasets ……..we'll cover this at the end of the exercise).

So, we have observed data (cells F4:F11) and expected data given the parameter estimates (cells L4:L11), and we can now assess goodness of fit with a Chi-Square test. You might remember the Chi-Square formula from your introductory genetics class. It is written:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed datapoint and E is the corresponding expected value.

We'll start by subtracting the expected frequency from the observed frequency, squaring the result, and dividing by the expected frequency for <u>each</u> of the 8 histories. This is done in cells M4:M11. For instance, cell M4 has the equation =(F4-L4)^2/L4, and has the result 0.62. Our model said that we should expect 18.61 sites to have a 100 history, and 22 of them did; $(22-18.61)^2/18.61 = 0.62$. Adding these cell values across all 8 histories gives us the chi-square test statistic, which is computed in cell M12 (and reflected in cell N17).

What's normally done next is to evaluate the probability of getting the Chi-Square test statistic we did, given that the data came from a "null sampling distribution." Then, we find out where this p value falls on a null Chi-Square distribution. The p value associated with this chi-square test statistic is computed in cell O17 with the formula =CHIDIST(N17,L17). In our case, the p value is 0.5613, suggesting that there is no evidence for lack of fit. P values less than 0.05 traditionally have been considered as evidence of lack of fit,

though some people view p values < 0.1 as evidence of lack of fit. If the data don't fit, you can inspect cells M4:M11 to determine which histories were the "offending" histories (those cells with high values). Because each cell value is Chi-Square distributed with 1 degree of freedom, cell values where $(O-E)^2/E >$ 3.84 should be viewed with suspicion. There may be some plausible reasons (due to the biology of the species, or due to your sampling methods) that would explain the poor fit.

	G	H	I	J	K	L	M
2					Probability		
3	Parameter	Estimate?	Betas	MLE	of History	Expected	Chi-Square
4	p1	1	1.432238	0.80725	0.074	18.61	0.62
5	p2	1	0.92962	0.71700	0.273	68.30	0.32
6	p3	1	0.370786	0.59165	0.108	26.96	0.14
7	ψ	1	1.372416	0.79777	0.189	47.14	0.80
8					0.220	55.00	0.00
9					0.065	16.31	0.10
10					0.026	6.44	0.32
11					0.045	11.26	0.67
12					1	250	2.979239783

This might be a good time to discuss why the Pearson Chi-Square goodness of fit approach is often not useful, especially when the expected values are low. When the expected values are low, even an observation at a single site will cause the chi-square test statistic to sky-rocket. As a hypothetical example, suppose you ran a model where the observed number of sites with a history of 001 was 1, but the expected value is 0.01 sites. This single observation leads to a chi-square value of <u>at least</u> $(1-0.01)^2/0.01 = 81$, which would indicate a "significant" lack of fit…all because you found the species at a single site where it was not expected to occur! This problem is well described in the MARK helpfiles. Thus, when expected values are low (e.g., less than 2), the Pearson Chi-Square isn't useful for assessing fit. This is especially problematic for

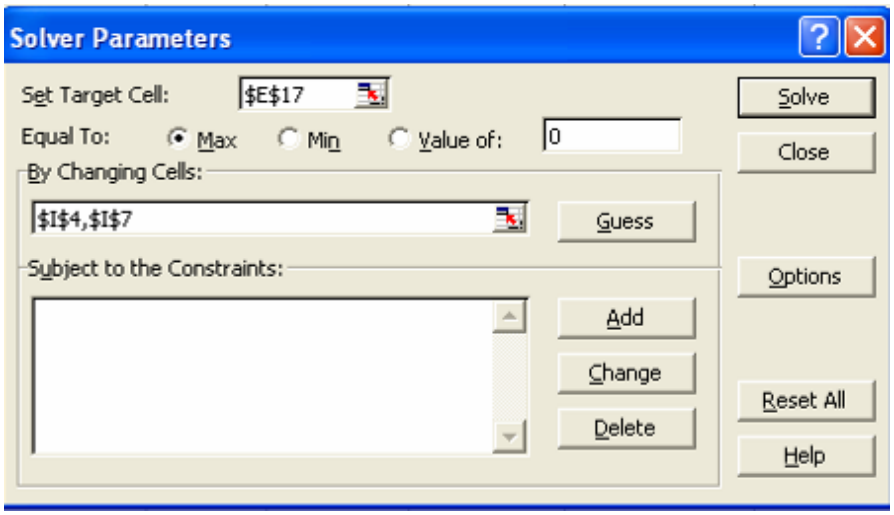
models with individual covariates because the expected values are almost always low. Fortunately, Darryl MacKenzie and Larissa Bailey figured out how to assess goodness of fit for occupancy models with covariates, and we'll discuss their approach in the next exercise.

COMPARING MODELS

You can compare different kinds of models to determine which model best fits the observed data by using model selection techniques (Burnham and Anderson 2002). For example, you might compare the last model (the most general model) with one in which p_1, p_2, and p_3 are constant (or model psi, p(.), where the dot indicates that p is forced to be constant over time). To run this model, you would simply enter 0 for p_2 and p_3 (since they are the same as p_1) in cells H5:H6, and you would enter =I4 in cells I5 and I6. In this way, you've constrained p_2 and p_3 to be equal to p_1, and thus you won't estimate them directly. So K = 2 for this model instead of 4.

	G	H	I
3	Parameter	Estimate?	Betas
4	p1	1	
5	p2	0	=I4
6	p3	0	=I4
7	ψ	1	

Now you can run Solver again. However, this time you'll maximize cell E17 by changing only cells I4 and I7.



Then press Solve, and keep the Solver solution. Now you can compare this reduced model (K = 2) to the full model (K = 4) by comparing their AICc scores. Here is the output we got:

	E	F	G	H	I	J	K	L	M	N	O
1	**Single-Species, Single-Season Model**										
2							Probability			**Saturated Model**	
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History	Expected	Chi-Square	Probability	Ln(Prob)
4	100	22	p1	1	0.858482	0.70234	0.050	12.46	7.30	0.088	-2.43041846
5	111	73	p2	0	0.858482	0.70234	0.278	69.39	0.19	0.292	-1.23100148
6	101	25	p3	0	0.858482	0.70234	0.118	29.41	0.66	0.1	-2.30258509
7	110	41	ψ	1	1.393356	0.80113	0.118	29.41	4.57	0.164	-1.80788885
8	000	55					0.220	55.00	0.00	0.22	-1.51412773
9	011	15					0.118	29.41	7.06	0.06	-2.81341072
10	001	5					0.050	12.46	4.47	0.02	-3.91202301
11	010	14					0.050	12.46	0.19	0.056	-2.88240359
12	# Sites =	250					1	250	24.43335294	Log L (sat)	-460.41234
13	# Histories =	8								-2 Log L (sat)	920.82467
14											
15	**OUTPUTS**										
16	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-473.15	946.2952919	2	950.30	950.34	920.8247	25.4706	6	4.245103459	24.4334	0.0004

You should notice that, as specified, $p_1 = p_2 = p_3$, and was estimated at 0.70234. Occupancy probability was 0.80113. This model has a much higher deviance than the first model we ran (25.4704), indicating it deviates from the saturated model more than the first model we ran. Also notice that the Pearson chi-square value is 24.433, with a p value of 0.004, indicating a lack of fit for this particular model. Examination of cells M4:M11 indicates that histories 100 and 011 were the most "offending" histories, with one having more histories than

expected and the other having fewer histories than expected. The AICc score for this model is 950.34, whereas the AICc score from the most general model (model phi(.)p(t)) was 931.95, a difference of 18.39 AIC units (ΔAICc = 18.39). In the AIC world, the first model is much more strongly supported than this one; there is virtually no support for the psi(.)p(.) model compared to the psi(.)p(t) model. This process is outlined fully in Burnham and Anderson 2002. We'll compare models with model selection approaches in the occupancy covariate exercise later on.

SIMULATING DATA

OK, we're almost done with the general, single-season occupancy model…hang in there! The next section describes two ways that you can simulate data. This is done for you in columns Q:W. The first thing you need to do is enter values for each parameter in the occupancy model in cells R4:V4. For example, enter 0.8 in cell R4 to simulate data where $p_1 = 0.8$, enter 0.7 in cell S4 to simulate data where $p_2 = 0.7$, and so on. The values that you see are the ones we used to simulate the data you just analyzed (the very first model you ran, psi(.)p(t), did a great job at estimating these parameters! That is, the estimates had low bias).

	Q	R	S	T	U	V	W
1	**Simulate Data**						
2							
3	Parameter	p1	p2	p3	ψ	N	
4	MLE	0.8	0.7	0.6	0.8	250	
5							
6							
7	**Summarized Stochastic Data:**				**Summarized Expected Data:**		
8	100	17	100 17;		100	19.2	100 19;
9	111	71	111 71;		111	67.2	111 67;
10	101	30	101 30;		101	28.8	101 29;
11	110	43	110 43;		110	44.8	110 45;
12	000	56	000 56;		000	54.8	000 55;
13	011	15	011 15;		011	16.8	011 17;
14	001	9	001 9;		001	7.2	001 7;
15	010	9	010 9;		010	11.2	010 11;
16		250				250	

Given these parameter values, there are two ways to simulate data. The first, and easiest, way is to simulate the data based on expected values. We've already touched on this subject when we calculated the Chi-Square test statistic. The histories for this method are written out in cells U8:U15. The number of sites expected to have each history are computed by entering a formula describing the history's probability, given the parameter values entered in cells R4:V4. For example, the number of sites expected to have a 100 history (cell V8) is $\psi*p_1*(1-p_2)*(1-p_3)*N$, which is 0.8*0.8*(1-0.7)*(1-0.6)*250 = 19.2 sites. Obtaining the expected number of sites for the other histories should be straight-forward to you by now: enter equations to generate the probability of each history given the parameters, and multiply the result by N to give you the expected number of sites. This was already done in cells V8:V15. You could copy the values in cells V8:V15, and then select cells F4:F11 and paste the new frequencies in. Of course, you'd have to round these expected numbers to 0

decimal places, and this could be done easily with a ROUND function. Simulating data based on expectation is a great way to evaluate model performance in terms of bias and precision (because you know beforehand what the results SHOULD be).

The second way to simulate data involves a stochastic (random) component, and this method is used in both MARK and PRESENCE for simulating "parametric" bootstrap data. Once again, the formula will reference the parameter values you entered in cells R4:V4. First, we set up a population with N individuals. In our case, we want to simulate data for 250 sites, so the sites are identified by number in cells Q20:Q269. Next, we identify a random ψ, p_1, p_2, and p_3 for each site, and this is done with a RAND function in cells R20:U269. (Each cell has the formula =RAND(), which generates a random number between 0 and 1).

	Q	R	S	T	U	V
19	**Site**	**rand() ψ**	**rand() p1**	**rand() p2**	**rand() p3**	**History**
20	1	0.70379984	0.37579	0.03366007	0.02379	111
21	2	0.78647418	0.52045	0.23830482	0.86203	110
22	3	0.66025016	0.42628	0.38845482	0.97714	110
23	4	0.24801038	0.86897	0.77132122	0.97551	000
24	5	0.33711909	0.01015	0.59995329	0.11561	111

Given these random numbers and the parameter estimates entered in cells R4:U4, we can determine each site's history (cells V20:V269). Let's walk through the equation. For site 1, the history is computed in cell V20 with the formula

=IF(AND(R20<$U$4,S20<$R$4),1,0)&IF(AND(R20<$U$4,T20<$S$4),1,0)&IF(AND(R20<$U$4,U20<$T$4),1,0). This formula is copied down for the remaining 249 sites.

This formula is pretty long and we could have broken it up into pieces, but it's not too bad once you understand that it really is 3 separate functions, one for each sampling period, which are combined together with a concatenation (&) symbol. The first function is =IF(AND(R20<$U$4,S20<$R$4),1,0), and determines whether the species was detected or not during sampling session 1. The equation basically says, IF cell R20 (the random ψ) is less than cell U4 (the specified ψ) AND cell S20 (the random p_1) is less than cell R4 (the specified p_1), then the species was detected and a 1 will result; otherwise the species was not detected in survey 1 and a 0 will result. The second function is IF(AND(R20<$U$4,T20<$S$4),1,0) and pertains to the second survey outcome. IF cell R20 (the random ψ) is less than cell U4 (the specified ψ), AND cell T20 (the random p_2) is less than cell S4 (the specified p_2), then the species was detected on the second survey and a 1 is recorded; otherwise the species was not detected on the second survey and a 0 is recorded. The third function is IF(AND(R20<$U$4,U20<$T$4),1,0) and generates the result of the third survey. IF cell R20 (the random ψ) is less than cell U4 (the specified ψ), AND cell U20 (the random p_3) is less than cell T4 (the specified p_3), then the species was detected on the third survey and a 1 is recorded; otherwise the species was not detected on the third survey and a 0 is recorded.

Each of the three IF functions are evaluated, resulting in a 0 or 1. The three numbers are then concatenated to give the final history. Hopefully this makes sense to you. Every time you press F9, the calculate key, Excel generates new random numbers, and hence new, simulated histories for each site.

The stochastic data are summarized in cells Q8:S15. The histories are listed in cells Q8:Q15, and the number of sites with each history is counted with a COUNTIF function in cells R8:R15. These data can be copied and pasted into cells F4:F11 for analysis. Press F9, and you should see a new, summarized data set with each button press. If you were to run the analysis in MARK, the MARK input file is generated for you in cells S8:S15.

	Q	R	S
7	**Summarized Stochastic Data:**		
8	100	20	100 20;
9	111	68	111 68;
10	101	23	101 23;
11	110	53	110 53;
12	000	42	000 42;
13	011	25	011 25;
14	001	7	001 7;
15	010	12	010 12;
16		250	

If you plan to do your analysis in PRESENCE, the input file is created for you in columns Y:AC.

	Y	Z	AA	AB	AC
1	PRESENCE INPUT				
2	Site	History	1-1	1-2	1-3
3	1	100	1	0	0
4	2	100	1	0	0
5	3	100	1	0	0
6	4	100	1	0	0
7	5	100	1	0	0

SINGLE SEASON OCCUPANCY MODELS ANALYSIS IN PRESENCE

OBJECTIVES

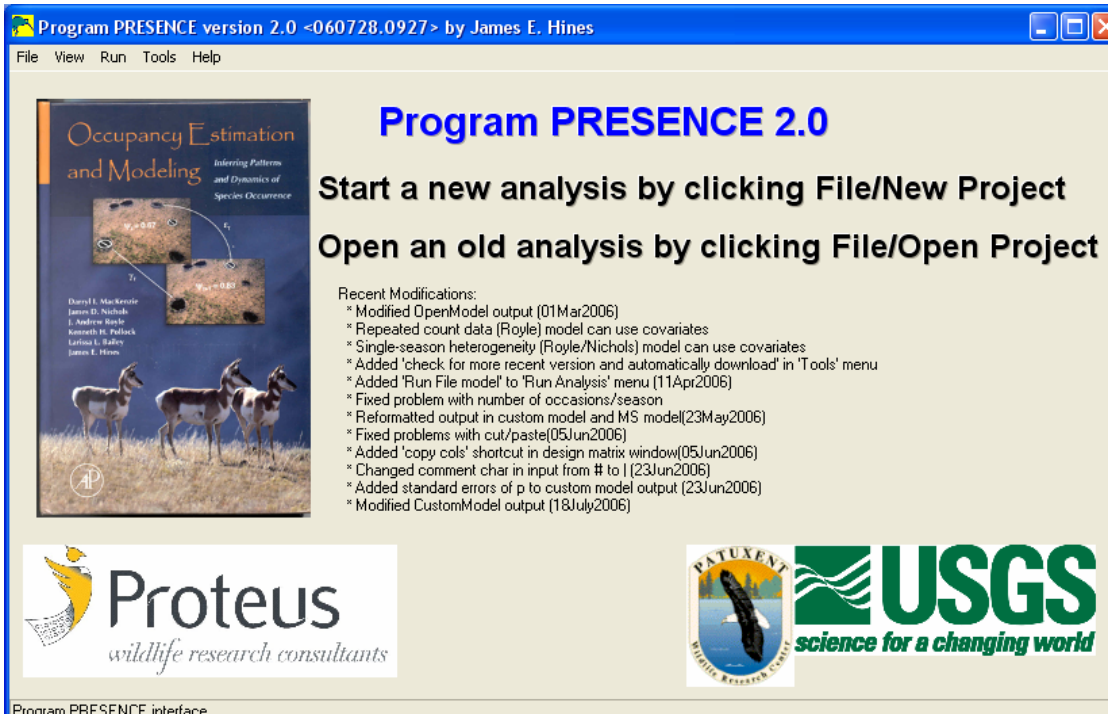
- To become familiar with PROGRAM PRESENCE
- To run the single-season occupancy model in PRESENCE
- To understand the PRESENCE output

DOWNLOADING PRESENCE

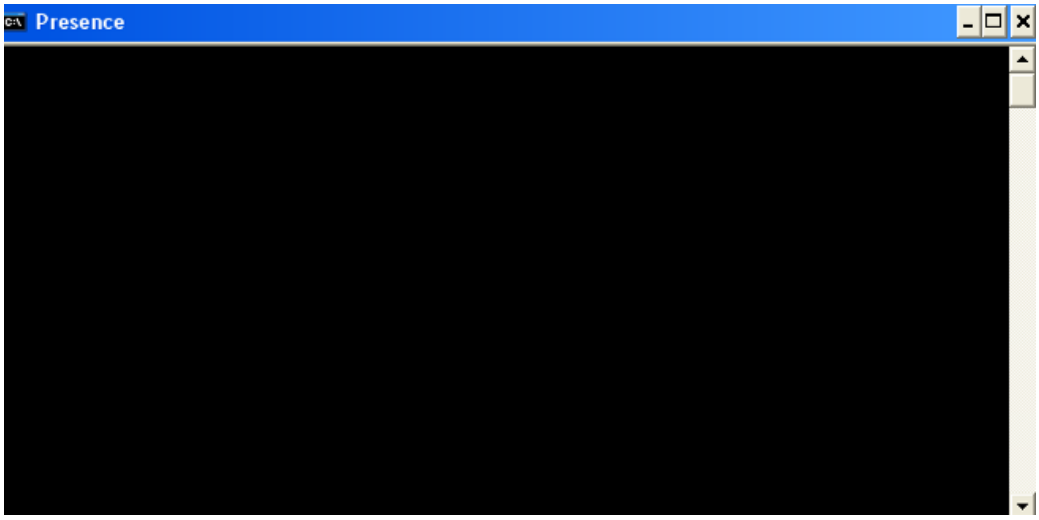
PRESENCE is a software program that runs many kinds of occupancy analyses. The first version was developed by Darryl MacKenzie, and the second version was written by Jim Hines. This program is free and can be downloaded at http://www.mbr-pwrc.usgs.gov/software/doc/presence/presence.html. In this exercise, we will be analyzing the data we explored in the previous (spreadsheet) chapter.

GETTING STARTED

When you open PRESENCE for the first time, you'll see the following screen:

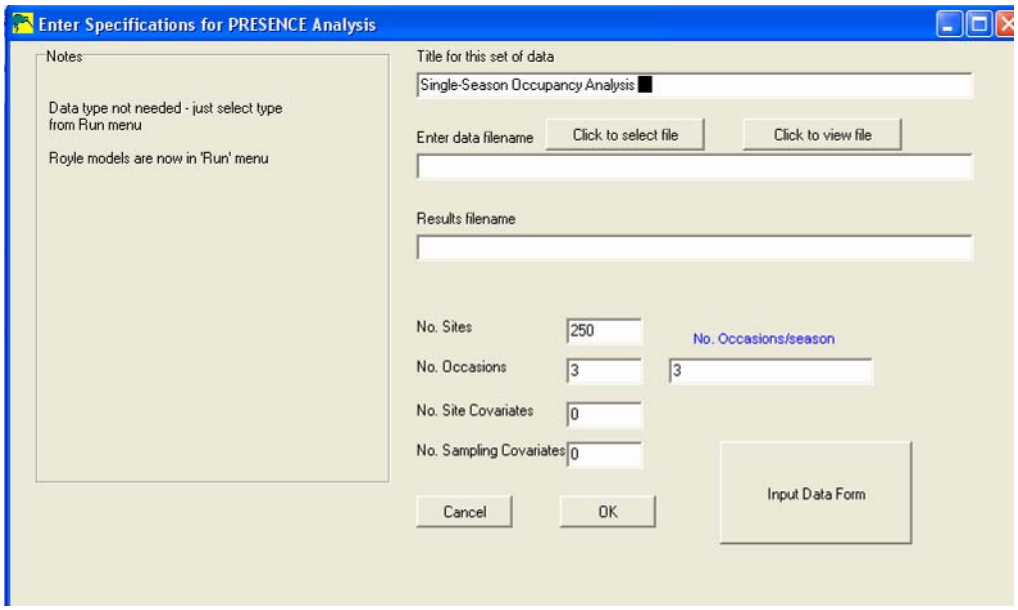


A second screen also appears, and is the engine. It looks like this, and should be kept open (though you can minimize it).

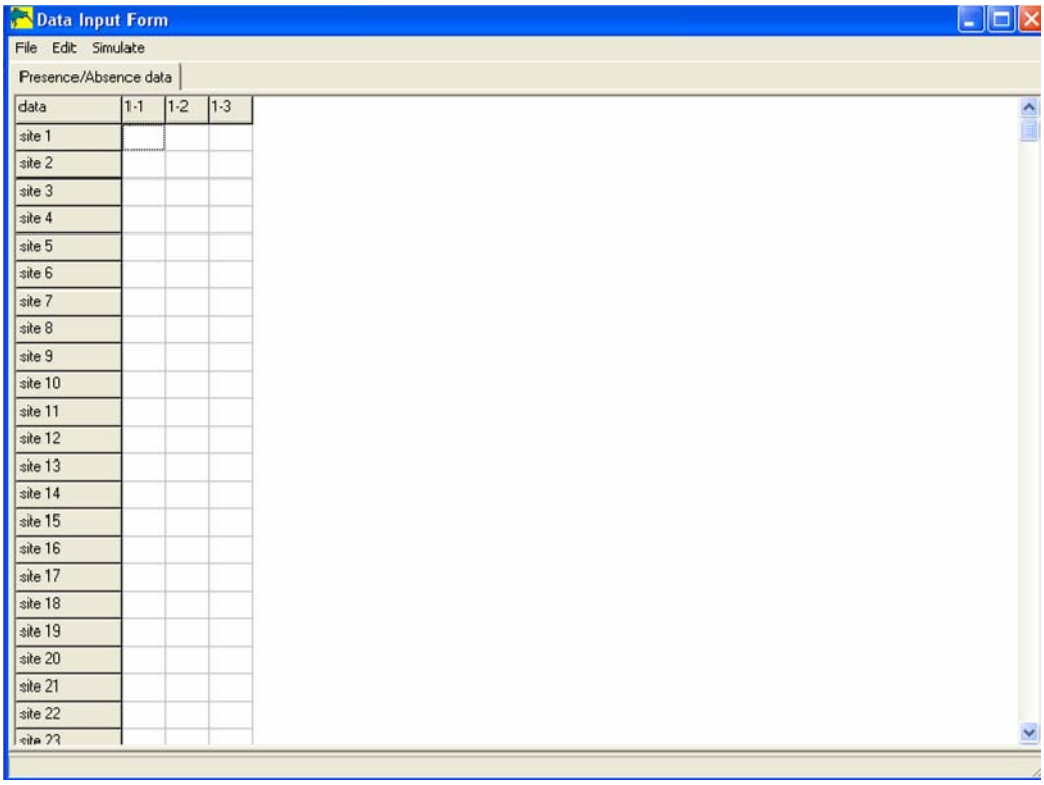


THE PRESENCE INPUT FILE

To begin a new analysis, simply click on File | New and you'll see the following screen, which is a form called the **Specifications for PRESENCE Analysis Form**:



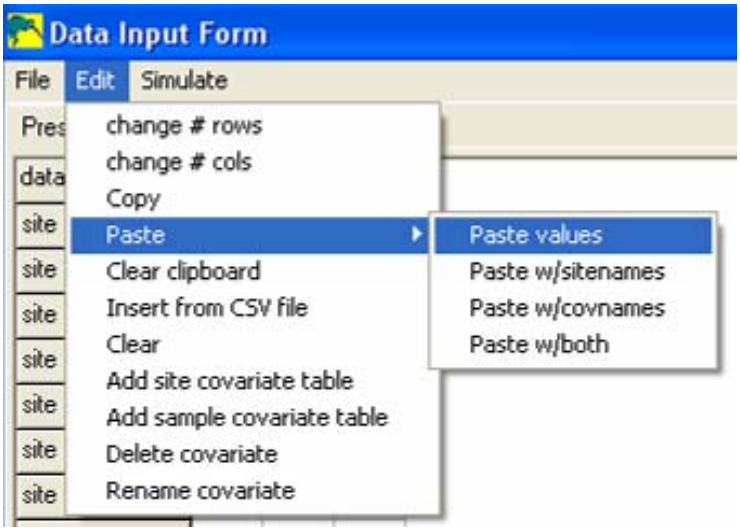
Enter a title for the analysis (e.g., Single-season occupancy analysis), and then provide some basic information in the text boxes at the bottom. Enter 250 for No. Sites, and enter 3 for No. Occasions. Then enter 3 for No. Occasions/Season. In this analysis, there are no site covariates or sampling covariates, so keep those as 0. Then click on the button labeled "Input Data Form", and the following screen will appear:



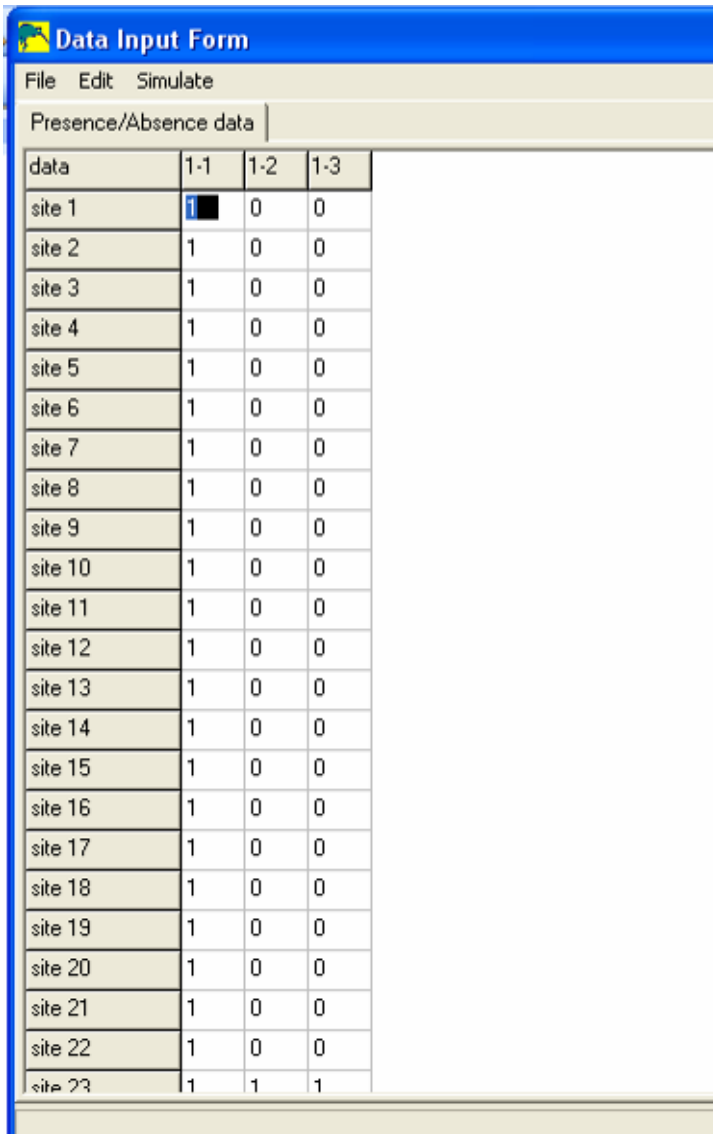
The sites are listed individually, and you'll paste in the raw data that is provided in columns AA:AC your spreadsheet (labeled Occupancy). The histories for the first 10 sites on the spreadsheet are shown below. These are computed values (click on any entry in column Z:AC and you'll see the underlying functions).

	Y	Z	AA	AB	AC
1	PRESENCE INPUT				
2	Site	History	1-1	1-2	1-3
3	1	100	1	0	0
4	2	100	1	0	0
5	3	100	1	0	0
6	4	100	1	0	0
7	5	100	1	0	0
8	6	100	1	0	0
9	7	100	1	0	0
10	8	100	1	0	0
11	9	100	1	0	0
12	10	100	1	0	0

Select cells AA3:AC252 and copy them. Then, click on the PRESENCE data input form and go to Edit | Paste | Paste values.

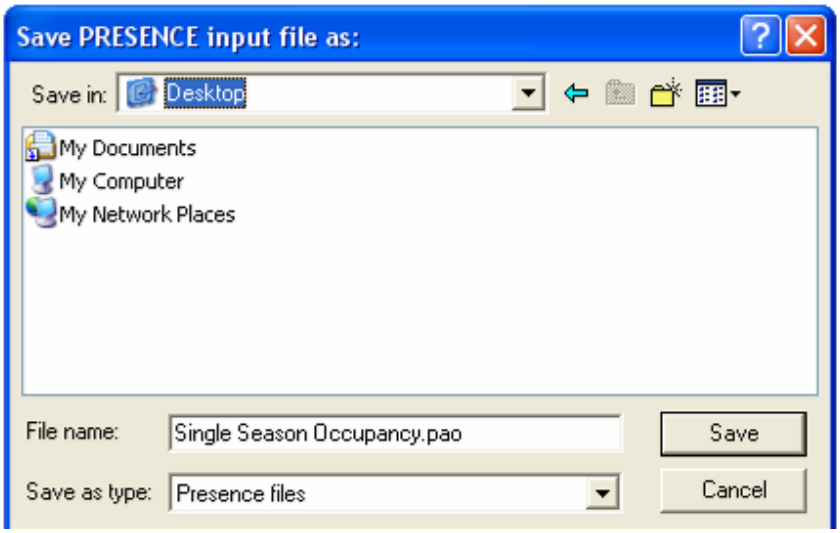


You could also paste in site names and covariates, and we'll do that in a different exercise. Your data input file should look like this (below). Make sure that the histories from all 250 sites are included.



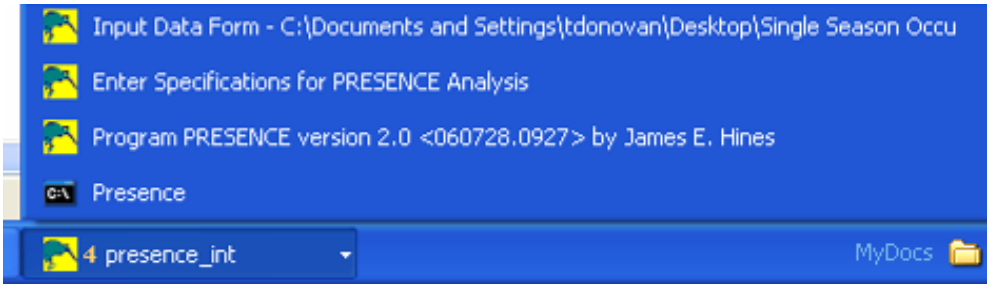
data	1-1	1-2	1-3
site 1	1	0	0
site 2	1	0	0
site 3	1	0	0
site 4	1	0	0
site 5	1	0	0
site 6	1	0	0
site 7	1	0	0
site 8	1	0	0
site 9	1	0	0
site 10	1	0	0
site 11	1	0	0
site 12	1	0	0
site 13	1	0	0
site 14	1	0	0
site 15	1	0	0
site 16	1	0	0
site 17	1	0	0
site 18	1	0	0
site 19	1	0	0
site 20	1	0	0
site 21	1	0	0
site 22	1	0	0
site 23	1	1	1

The next step is to SAVE this file as the PRESENCE input file. Go to File | Save As, and you'll see the following input screen:

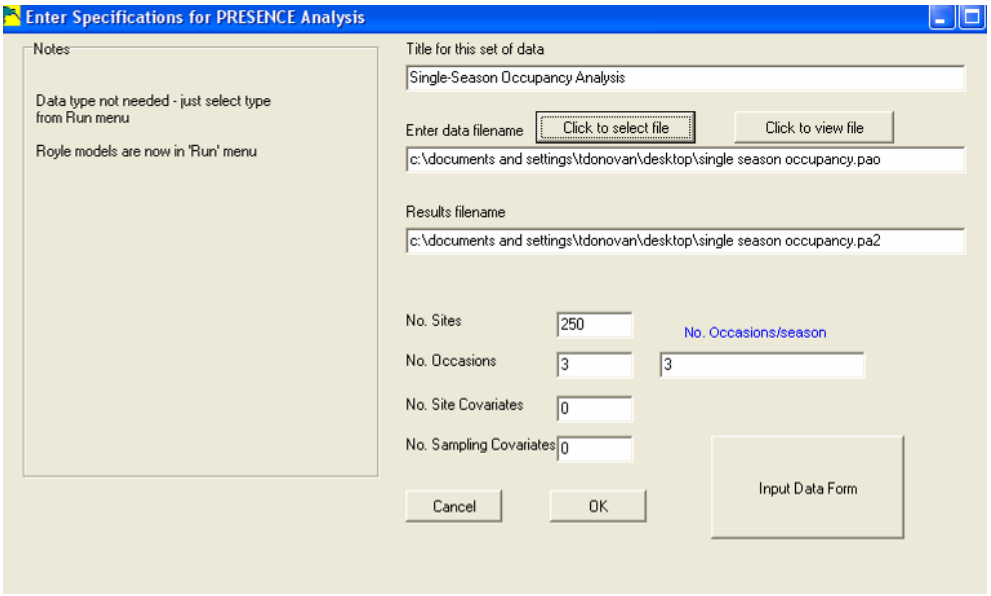


Enter a title for your file, such as Single Season Occupancy, and keep the pao extension. The PAO extension in PRESENCE stands for Proportion of Area Occupied, and is the input file that PRESENCE recognizes. Navigate to a place on your computer where you keep your PRESENCE files (we put it on the desktop). Click SAVE.

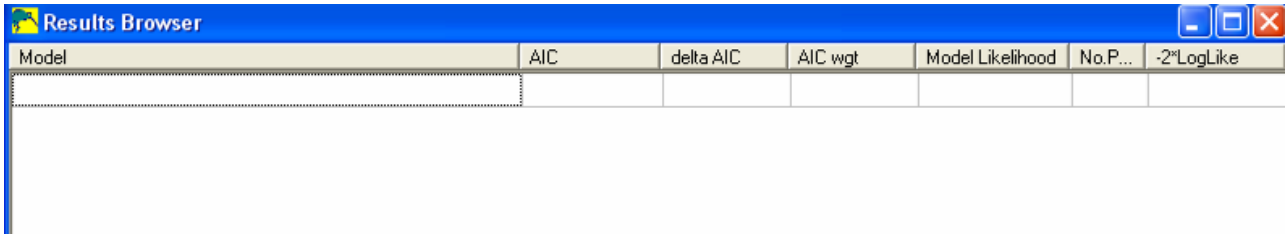
Notice that there are now 4 windows in PRESENCE that are open.



Click again on the Specifications for PRESENCE Analysis form. Now we need to tell PRESENCE where to find the data file we just created. Click on the button labeled "Click to Select File" and navigate to the data file you just created and double click the file.



You should see that PRESENCE automatically creates a Results filename, with the same name as the data file name but with the extension .pa2. You've finished entering your Specifications. Click OK. PRESENCE then spawns the Results Browser, and your screen should look like this:

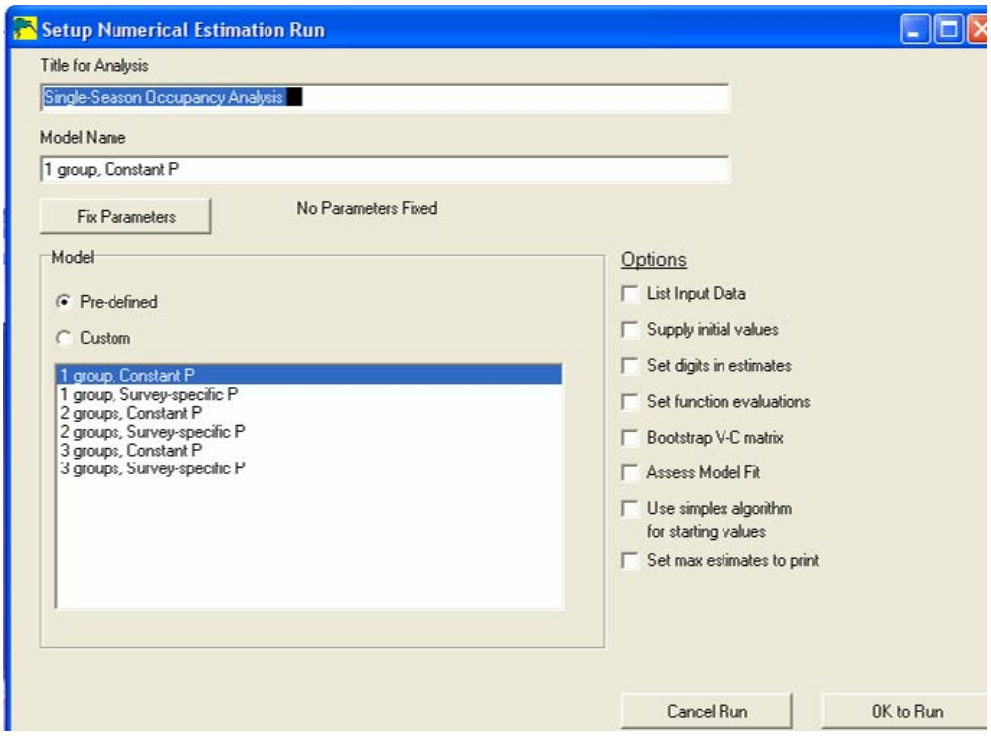


Now you are ready to run an analysis in PRESENCE.

MODEL PSI,P(T)

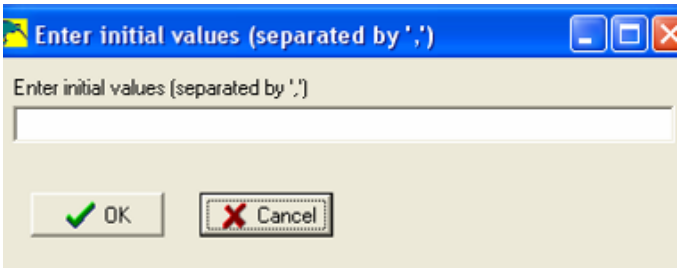
PROGRAM PRESENCE has several analytical options, and the first thing you need to do is specify that you are running a single-species, single-season analysis. Open the main window, and select Run | Analysis: single-season.

PRESENCE will then provide the following input page, which is labeled **Setup Numerical Estimation Run:**



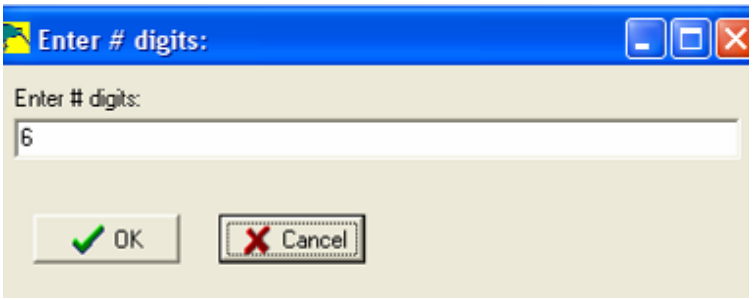
The title shown is the one you provided earlier. Notice that this input form has several options. There is a text box named **Model Name**, where you provide the name of a model you will be running. We will be running model psi, p(t) shortly. Under the model name text box, there is a group box labeled Model, and here you select whether you want to run pre-defined models or custom models. Note that by default, pre-defined is selected and six pre-defined models are labeled: (1) 1 group, constant p, (2) 1 group, Survey-specific p, (3) 2 group2, constant p, (4) 2 group2, Survey-specific p, (5) 3 groups, constant p, (6) 3 groups, Survey-specific p. In the spreadsheet, we ran only the first two options, one where p was forced to be constant and one where we estimated p1, p2, and p3 individually. Options 3-6 are mixture models and we will explore them in a separate exercise. In this exercise, we will run pre-defined models 1 and 2. Later we will be using the Custom option so that we can include covariates.

On the right side of the Setup Numerical Estimation Run window are several check-box options. In the right hand portion of the screen, you can enter various options for running the analysis. The box labeled **List Input Data** will list your raw data in the PRESENCE output file. The box labeled **Supply Initial Values** allows you to enter starting estimates of the parameters. This is sometimes useful to help PRESENCE find the multinomial maximimum log likelihood most efficiently. Though we don't need to do this for this exercise, you might keep this in mind for other analyses. If you select this option, the following dialogue box will appear before the model is run.



Enter first an initial estimate for ψ, then enter estimates for p_1, p_2, and p_3 respectively.

The option, **Set Digits in Estimates**, allows you to control how many digits will appear in the parameter estimates. If you select this option, the following dialogue box will appear:



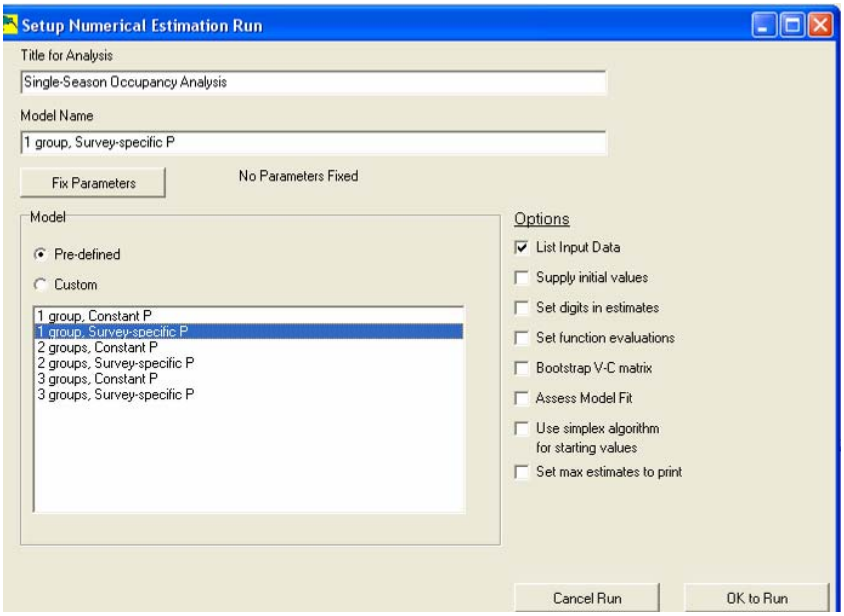
The option, **Set Function Evaluations**, allows you to control the maximum number of functions that PRESENCE will evaluate before returning an answer. The default is 2000.

The option **Bootstrap VC Matrix** allows you to obtain variance-covariance estimates through bootstrapping, and the option **Assess Model Fit** runs the MacKenzie and Bailey Goodness of Fit test. We'll use this option in a later exercise. The box labeled **Use Simplex Algorithm for Starting Values**

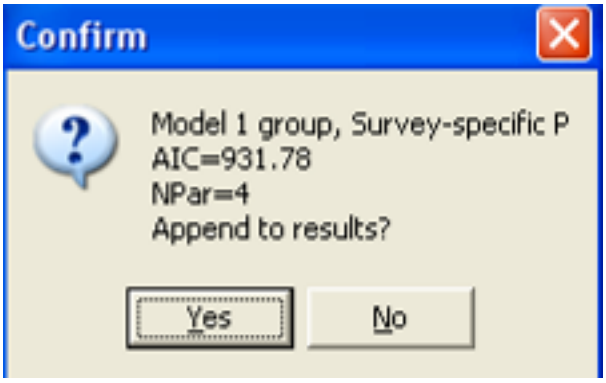
indicates that you want PRESENCE to use an alternative method for finding the maximum log likelihood. The program MARK provides a similar option, and here is how Gary White explains this method in the MARK helpfiles:

"The second method of optimization is simulated annealing. Simulated annealing is a global optimization method that distinguishes between different local optima. Starting from an initial point, the algorithm takes a step and the function is evaluated. When minimizing a function, any downhill step is accepted and the process repeats from this new point. An uphill step may be accepted. Thus, simulated annealing can escape from local optima. This uphill decision is made by the Metropolis criteria. As the optimization process proceeds, the length of the steps decline and the algorithm closes in on the global optimum. Since the algorithm makes very few assumptions regarding the function to be optimized, it is quite robust with respect to non-quadratic surfaces. Simulated annealing can be used as a local optimizer for difficult functions."

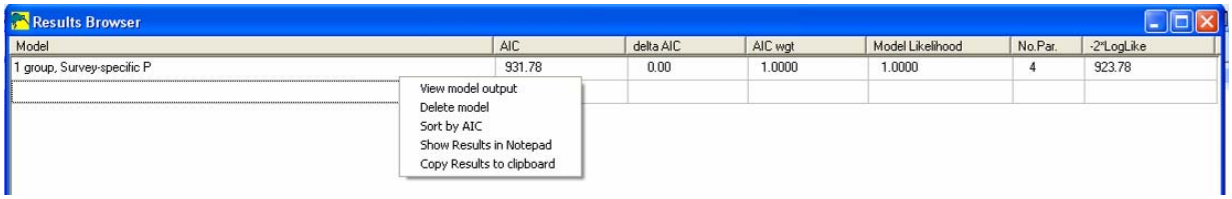
We'll start by running the pre-defined model, 1 group, Survey-specific p. By selecting that option, PRESENCE automatically provides the appropriate model name. Once you have checked the boxes you want, click OK to Run.



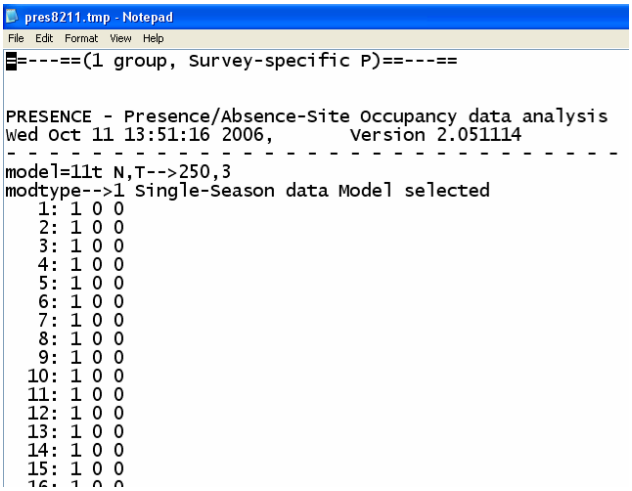
PRESENCE runs the analysis, and then provides a dialogue box asking whether you want the results pasted into the Results Browser:



Click Yes, and the results will be added:



If for some reason the Results Browser does not appear, click on View | Results. The Results Browser provides us with the model's AIC, delta AIC, AIC weight, the number of parameter, and the $-2\text{Log}_e\text{L}$. Now, we really want to see ALL of the outputs. In PRESENCE, just right-click on the model name in the Results Browser, and you'll see several options. Click on the option, View Model Output, and PRESENCE provides the full output in NotePad:



There is a lot of information in the output. First you'll notice that the model's name is on the top line of the output, followed by information about the date of the analysis and version number. Because we selected the "List Data" option, the raw data is provided. Scroll down further until you get to the meat of the output:

pres8211.tmp - Notepad

File Edit Format View Help

```
 249: 0 1 0
 250: 0 1 0
NSi-->0
NSa-->0

Site Covariates:

No Site covariates

 Sample Covariates:- - - - - - - - - - - - - - - - - - - - - - - - - - -
 Single-Season Occupancy Analysis
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
modtype=1 N=250 T=3 Groups=1 bootstraps=0

==>0


Predefined Model: Detection probabilities are time-specific

Number of groups               = 1
Number of sites                = 250
Number of sampling occasions   = 3
Number of missing observations = 0

Number of parameters           = 4
-2log(likelihood)              = 923.784168
AIC                            = 931.784168
Naive estimate                 = 0.780000

Proportion of sites occupied    (Psi)   = 0.7978 (0.027105)
Probability of group membership (Theta) = 1.0000
Detection probabilities         (p):
  grp   srvy      p            se(p)
  ---   ----   ---------    -----------
   1      1    0.807250     ( 0.029523)
   1      2    0.716998     ( 0.033008)
   1      3    0.591649     ( 0.035504)

Variance-Covariance Matrix
   psi     p1(G1)  p2(G1)  p3(G1)
 0.0007 -0.0001 -0.0001 -0.0001
-0.0001  0.0009  0.0001  0.0001
-0.0001  0.0001  0.0011  0.0001
-0.0001  0.0001  0.0001  0.0013
-----------------------------------------------------
CPU time: 0.0 seconds
```

The number of groups in this analysis is 1 (no mixtures), and there were 250 sites. There were 3 sampling occasions and 0 missing observations. The number of parameters for this model is 4 (ψ, p_1, p_2, p_3 are each estimated individually), and the -2Log$_e$L is 923.78. The AIC is 931.78. The **naïve estimate** is 0.78. What's this? It is the proportion of sites where the animal was detected at least once, and is computed in cell F14 on the spreadsheet. The Proportion of sites occupied is ψ, and is 0.7978 with a standard error of 0.027105 (listed in

parentheses). The probability of group membership (theta) is 1.00 and this parameter is important in mixture models – but since we didn't include mixtures in this analysis all sites belong to group 1. The detection probabilities are listed in a table: p_1 = 0.807250 (se = 0.02953), p_2 = 0.716998 (se = 0.033008), p_3 = 0.591649 (se = 0.03550). For comparison, the spreadsheet results from this model are provided and match PRESENCE.

	E	F	G	H	I	J	K
2							Probability
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History
4	100	22	p1	1	1.432238014	0.80725	0.074
5	111	73	p2	1	0.929620335	0.71700	0.273
6	101	25	p3	1	0.37078615	0.59165	0.108
7	110	41	ψ	1	1.372416073	0.79777	0.189
8	000	55					0.220
9	011	15					0.065
10	001	5					0.026
11	010	14					0.045
12	# Sites =	250					1
13	# Histories =	8					

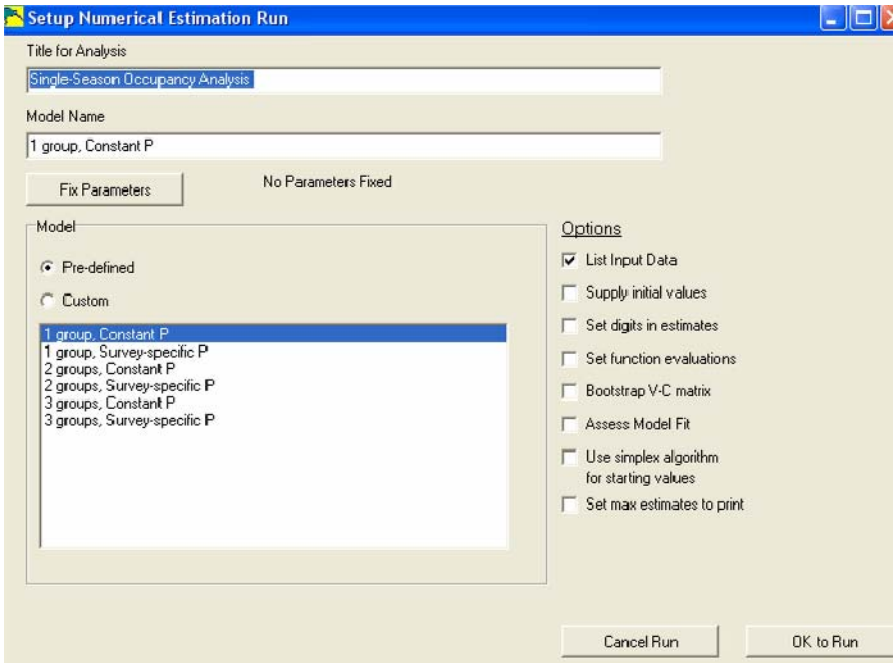
	E	F	G	H	I	J	K	L	M	N	O
15	OUTPUTS										
16	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-461.89	923.7841681	4	931.78	931.95	920.8247	2.9595	4	0.73987425	2.9792	0.5613

Keep in mind that the spreadsheet does not estimate the variances or the variance covariance matrix, and these are absolutely essential when interpreting your results.

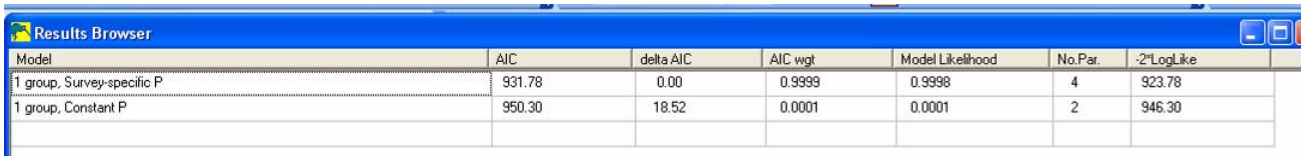
MODEL PSI,P(.)

Now let's run our second model, model PSI, P(.), in which we force $p_1 = p_2 = p_3$. Recall that in the spreadsheet, we ran this model by forcing the betas for p_2 and p_3 to equal p_1, and then solved. To run this model in PRESENCE, go to Run |

Analysis | Single-Season, and then select the pre-defined model 1 group, Constant P in the Setup Window.



Select the options you desire, then press OK to Run and append the results to the Results Browser:



Model	AIC	delta AIC	AIC wgt	Model Likelihood	No.Par.	-2*LogLike
1 group, Survey-specific P	931.78	0.00	0.9999	0.9998	4	923.78
1 group, Constant P	950.30	18.52	0.0001	0.0001	2	946.30

Here is the output from the spreadsheet model, where the AIC score is given in cell H17, K is given in cell G17, and the $-2\text{Log}_e\text{L}$ is given in cell F17.

	E	F	G	H	I	J	K	L	M	N	O
1	**Single-Species, Single-Season Model**										
2							Probability			**Saturated Model**	
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History	Expected	Chi-Square	Probability	Ln(Prob)
4	100	22	p1	1	0.858482	0.70234	0.050	12.46	7.30	0.088	-2.43041846
5	111	73	p2	0	0.858482	0.70234	0.278	69.39	0.19	0.292	-1.23100148
6	101	25	p3	0	0.858482	0.70234	0.118	29.41	0.66	0.1	-2.30258509
7	110	41	ψ	1	1.393356	0.80113	0.118	29.41	4.57	0.164	-1.80788885
8	000	55					0.220	55.00	0.00	0.22	-1.51412773
9	011	15					0.118	29.41	7.06	0.06	-2.81341072
10	001	5					0.050	12.46	4.47	0.02	-3.91202301
11	010	14					0.050	12.46	0.19	0.056	-2.88240359
12	# Sites =	250					1	250	24.43335294	Log L (sat)	-460.41234
13	# Histories =	8								-2 Log L (sat)	920.82467
14											
15	**OUTPUTS**										
16	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-473.15	946.2952919	2	950.30	950.34	920.8247	25.4706	6	4.245103459	24.4334	0.0004

To view the rest of the model output in PRESENCE, just right click on the model name and work your way through the output. You should see that the spreadsheet found the same results. We don't recommend doing your analysis in the spreadsheet, but the spreadsheet hopefully provides an idea of how the output was computed.

pres7804.tmp - Notepad

File Edit Format View Help

```
 246: 0 1 0
 247: 0 1 0
 248: 0 1 0
 249: 0 1 0
 250: 0 1 0
NSi-->0
NSa-->0

Site Covariates:

No Site covariates

 Sample Covariates:- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Single-Season Occupancy Analysis
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
modtype=1 N=250 T=3 Groups=1 bootstraps=0

==>0


Predefined Model: Detection probabilities are NOT time-specific

Number of groups               = 1
Number of sites                = 250
Number of sampling occasions   = 3
Number of missing observations = 0

Number of parameters           = 2
-2log(likelihood)              = 946.295292
AIC                            = 950.295292
Naive estimate                 = 0.780000

Proportion of sites occupied    (Psi)   = 0.8011 (0.027001)
Probability of group membership (Theta) = 1.0000
Detection probabilities         (p):
  grp   srvy      p          se(p)
  ---   ----   ---------    -----------
   1      1    0.702343     ( 0.013776)

Variance-Covariance Matrix
   psi    p(G1)
 0.0007 -0.0000
-0.0000  0.0002
-------------------------------------------------------
CPU time: 0.0 seconds
```

SUMMARY

The purpose of this exercise was to introduce you to the program, PRESENCE. There are many options in this program, and we'll work through them as we progress through the book.

SINGLE SEASON OCCUPANCY MODELS ANALYSIS IN PROGRAM MARK

OBJECTIVES

- To become familiar with PROGRAM MARK
- To run the single-season occupancy model in MARK
- To understand MARK's output

DOWNLOADING MARK

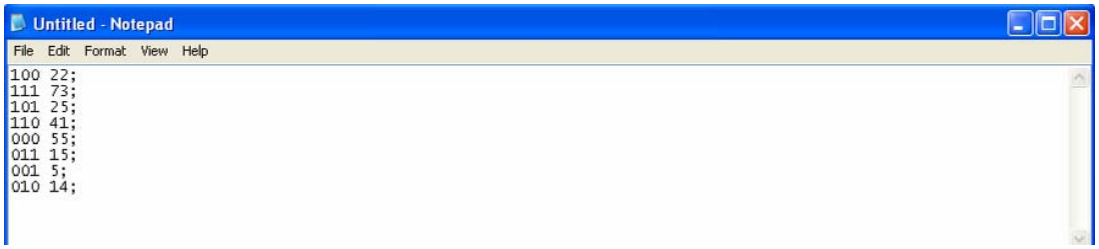
MARK is a program developed by Gary White at Colorado State University, and is freely distributed. The program can be downloaded from the MARK homepage at http://www.cnr.colostate.edu/~gwhite/mark/mark.htm.
You'll want to make sure that you are running the most current version of MARK, as Gary continuously updates the program.

GETTING STARTED: CREATING AN INPUT FILE

Hopefully you've worked through the spreadsheet Occupancy Models by now. Before you begin, you'll need to create a MARK input file. We'll stick with the data we've been using (and won't simulate new data). Make sure your spreadsheet matches what's shown below:

	B	C	D	E	F
3	MARK			History	Frequency
4	100 22;			100	22
5	111 73;			111	73
6	101 25;			101	25
7	110 41;			110	41
8	000 55;			000	55
9	011 15;			011	15
10	001 5;			001	5
11	010 14;			010	14
12				# Sites =	250
13				# Histories =	8

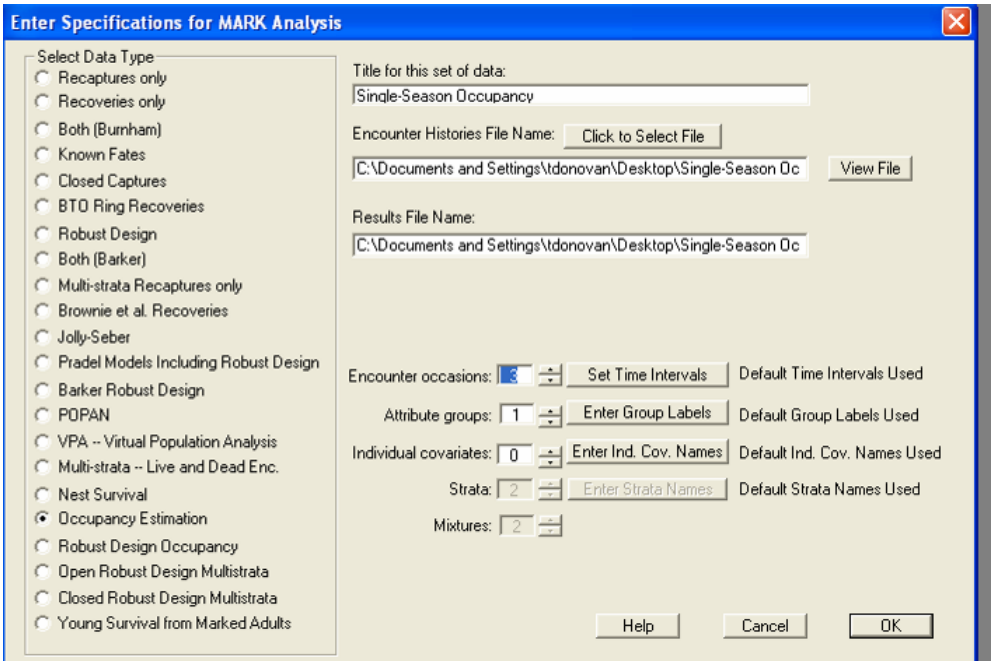
To analyze these data in MARK, select cells B4:B11, and copy them to your clipboard. Then open NotePad, and paste the values in. Note the MARK's input requires that you first enter the history, followed by a space, followed by the frequency, followed by a semicolon.



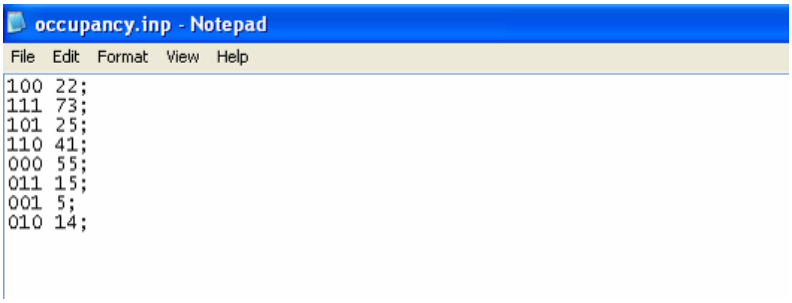
Save this file as "Single-Season Occupancy Input.inp" (include the quotes) onto your desktop (or somewhere where you know where to find it). This is the file you'll import it into MARK.

Open MARK and go to File | New. A new window will appear. Click on the Occupancy Estimation button on the left side of the screen. Type in a title for your analysis, e.g., Single-Season Occupancy. Then click on the button "Click to Select File" and then browse to your .inp file.

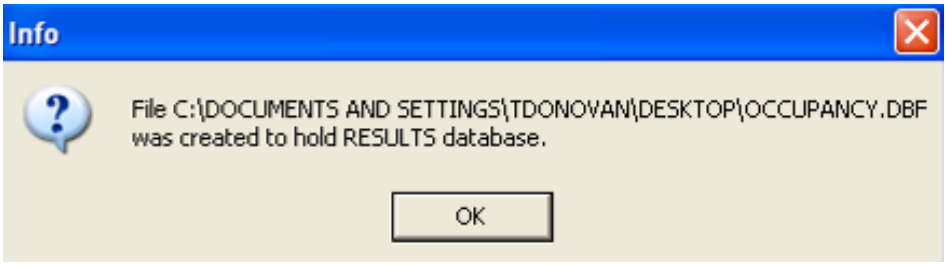
In this example, we had 3 encounter occasions, so set this to 3. There is only 1 group in this first analysis, and there are no covariates.



Click on the button labeled "View File" to make sure you uploaded the correct file:



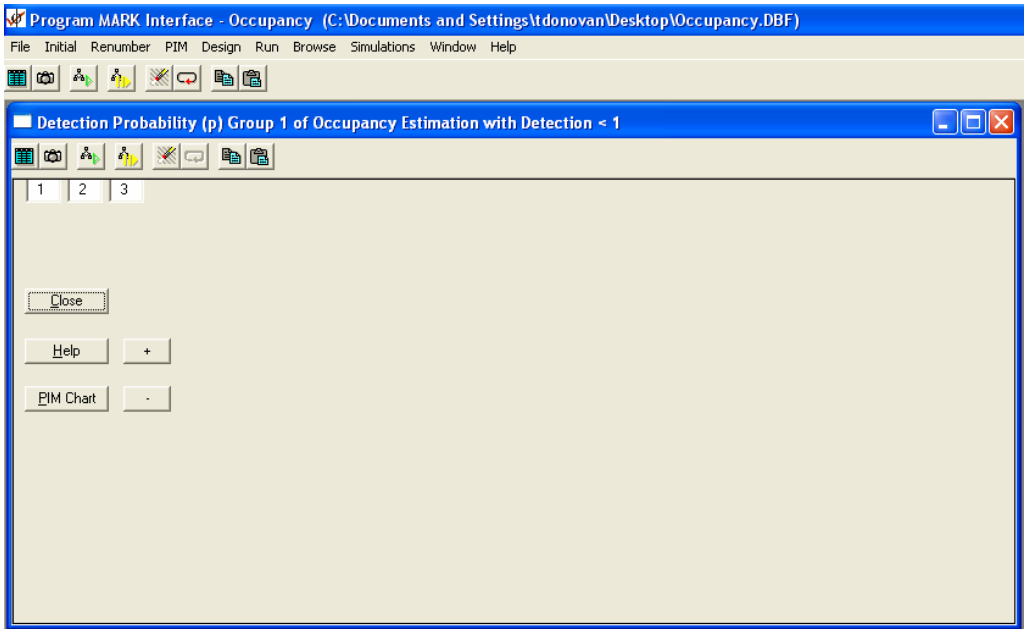
Once you're sure that everything is correct, click the button labeled OK and MARK will tell you that it will create a DBF file to hold the results. Click OK.



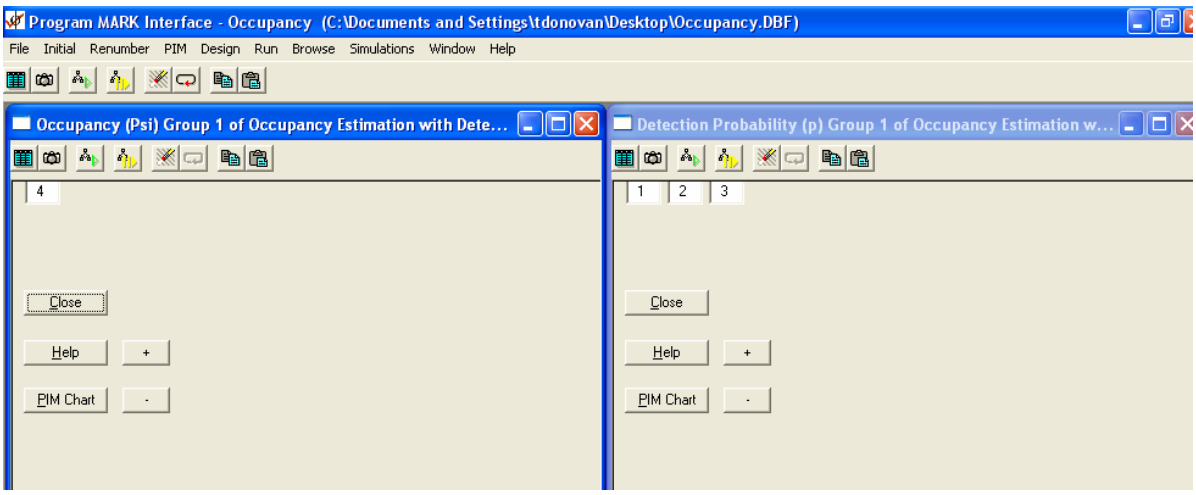
MARK PIM (PARAMETER INDEX MATRIX)

The Parameter Index Matrix will then appear. If you're comfortable with the PIM, great! If not, you'll become familiar with it soon enough. Basically, the PIM identifies which parameters you'll be estimating for a particular model, and numbers the parameters. For instance, the PIM below shows the parameters associated with Detection Probability (p). There are three parameters to be estimated: p_1, p_2, and p_3 , and by entering the numbers 1, 2, and 3 in the PIM boxes, you're telling MARK that you want a separate estimate for each, and that p_1 will be called Parameter 1, p_2 will be called Parameter 2, and p_3 will be called Parameter 3. MARK will report these Parameter indices, as well as the name of the parameter, in the model output. If you've finished the spreadsheet exercise, the PIMS are akin to the following section of the spreadsheet, where you indicate which parameters you are interested in estimating separately, and which ones you'll constrain to be equal to others.

	G	H
3	Parameter	Estimate?
4	p1	1
5	p2	1
6	p3	1
7	φ	1



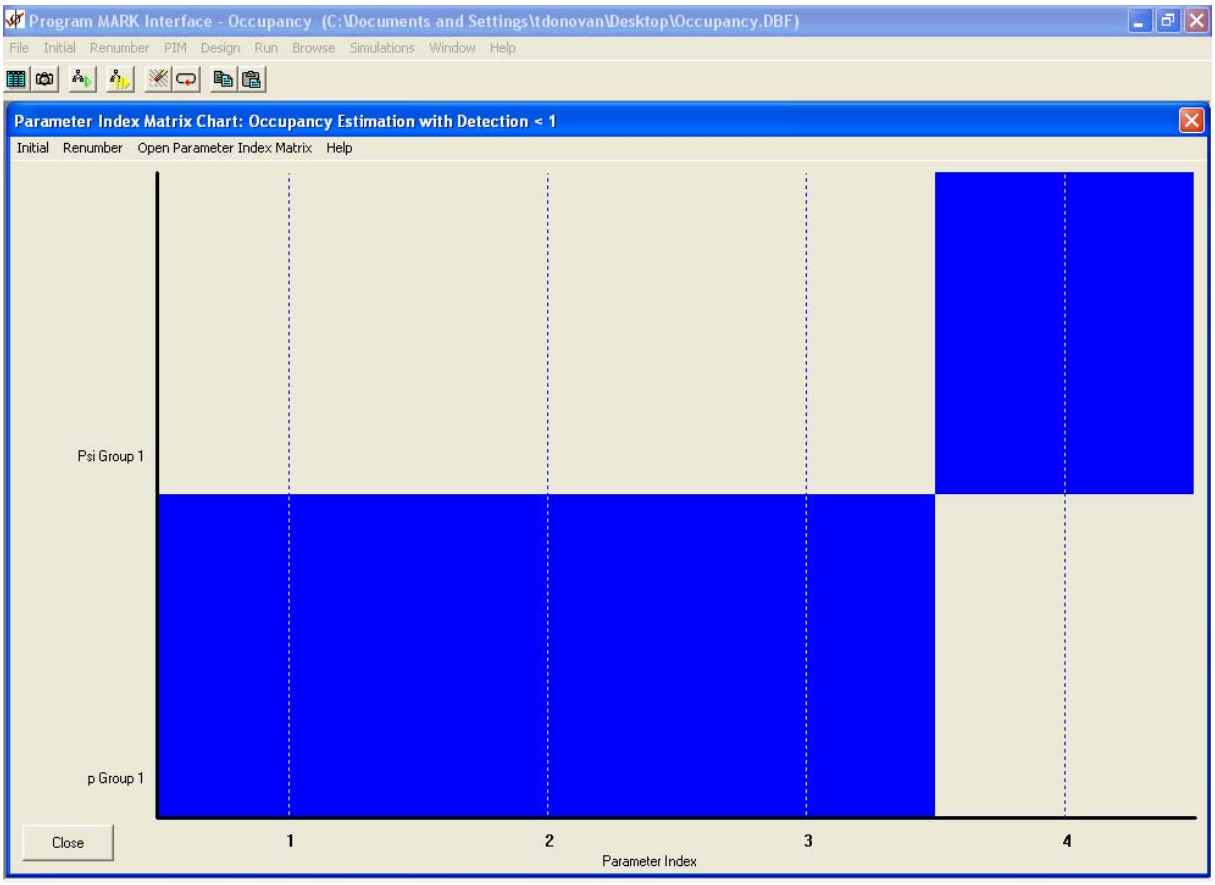
The other PIM chart is for the Occupancy rate, ψ. You might find it useful to look at both PIMs at the same time. Go to PIM | Open Parameter Index Matrix | and then select the Occupancy Group 1 PIM. Click OK. Now the PIM for occupancy is also open, and it shows a single parameter (labeled 4). It's useful to see both PIMs simultaneously, so go to Window | Tile, and you should see the following screen.



Now we can see there are 4 total parameters to be estimated. P_1 will be called Parameter 1, p_2 will be called Parameter 2, p_3 will be called Parameter 3, and ψ will be called Parameter 4. So MARK is currently set up to run the full model, similar to the one we did on the spreadsheet.

MARK PIM CHART

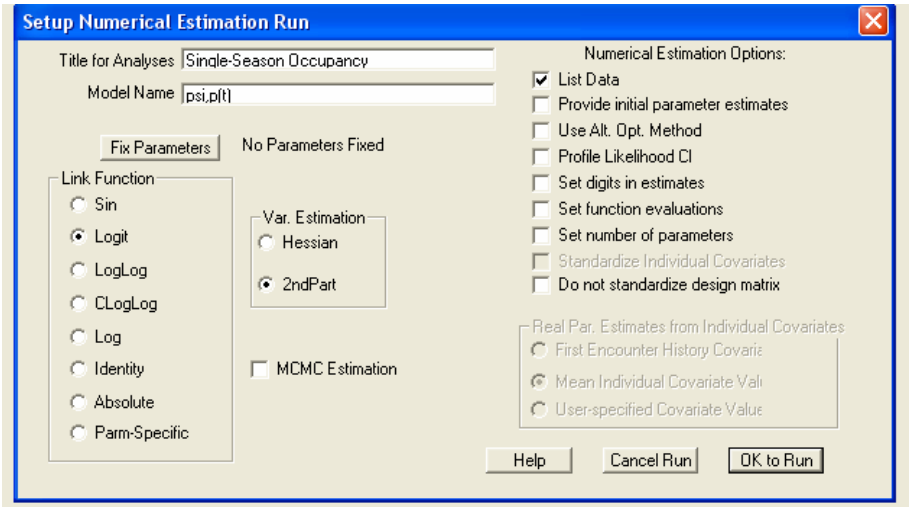
You can also view this as a PIM chart by clicking on the button PIM Chart, or by going to the toolbar PIM | Parameter Index Chart. The following screen will appear:



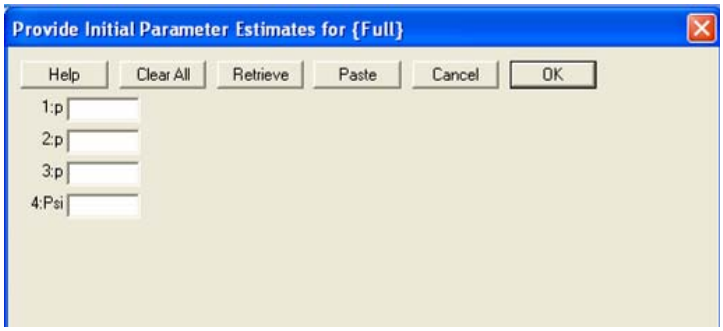
Note that the parameters 1-4 are listed on the bottom, and that there are 3 p estimates and 1 psi estimate (column labels are to the left of the image).

RUNNING THE FULL MODEL

You're ready to run this model. Go to Run | Current Model, and a new screen will appear. Type in a model name (e.g. Full, or Psi(.)P(t)). Under the section labeled "Link Function", select the Logit link radio button (because we used this link in the spreadsheet). The section labeled "Var. Estimation" indicates how you want MARK to estimate standard errors. The default is labeled "2ndPart", and estimates the variance-covariance matrix by estimating the second partial derivatives of the likelihood function. You should not use the Hessian option, as Gary White states "This approach is not reliable in that the resulting variance-covariance matrix is not particularly close to the true variance-covariance estimate." The box labeled MCMC Estimation should be checked if you're interested in running a Baysian Markov Chain Monte Carlo (MCMC), and is a Bayesian parameter estimation procedure that is most useful in MARK for estimating variance components.



In the right hand portion of the screen, you can enter various options for running the analysis. The box labeled "List Data" will list your raw data in the MARK output file. The box labeled "Provide Initial Parameter Estimates" allows you to enter starting estimates of the parameters. This is sometimes useful to help MARK find the multinomial maximimum log likelihood most efficiently. Though we don't need to do this for this exercise, you might keep this in mind for other analyses. If you select this option, the following dialogue box will appear before the model is run.



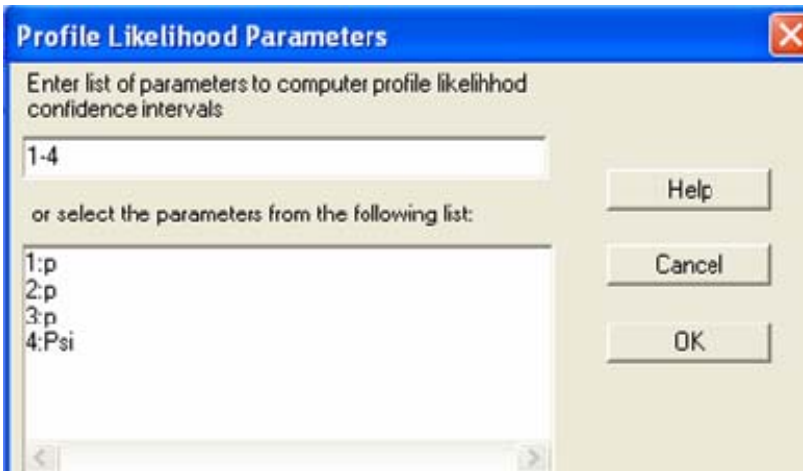
The box labeled "Use Alt. Opt. Method" indicates that you want MARK to use an alternative method for finding the maximum log likelihood. Here is how Gary White explains this method in the MARK helpfiles:

"The second method of optimization is simulated annealing. Simulated annealing is a global optimization method that distinguishes between different local optima. Starting from an initial point, the algorithm takes a step and the function is evaluated. When minimizing a function, any downhill step is accepted and the process repeats from this new point. An uphill step may be accepted. Thus, simulated annealing can escape from local optima. This uphill decision is made by the Metropolis criteria. As the optimization process proceeds, the length of the steps decline and the algorithm closes in on the global optimum. Since the algorithm makes very few assumptions regarding the function to be optimized, it is quite robust with respect to non-quadratic surfaces. Simulated annealing can be used as a local optimizer for difficult functions."

We won't use this method for the occupancy models, but if you might keep this option in mind (especially for multi-strata problems). Checking the box labeled "Profile Likelihood CI," indicates that you want MARK to output the profile likelihood confidence limits. Gary White explains that the

"Profile likelihood confidence intervals are based on the log-likelihood function. For a single parameter, likelihood theory shows that the 2 points 1.92 units down from the maximum of the log-likelihood function provide a 95% confidence interval. "

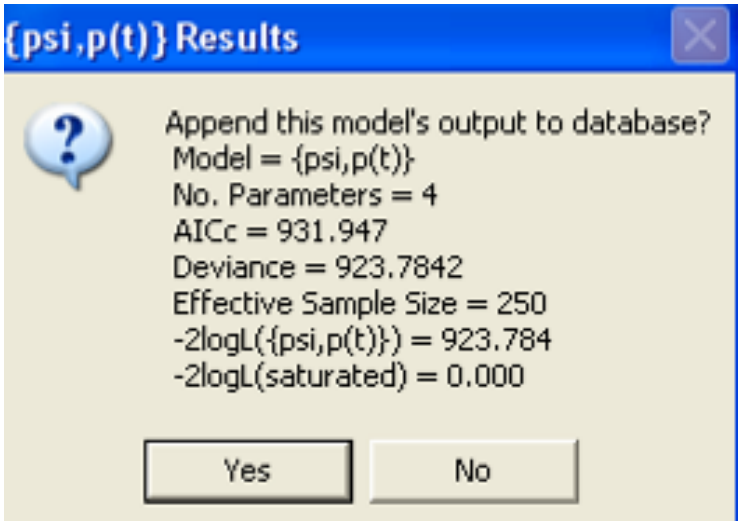
If you've checked the profile likelihood option, a dialogue box will appear before the model is run, and here you specify which parameters you want profile likelihood estimates. In this box, you enter the parameter indices (e.g., 1-4) rather than the parameter name.



The last three check boxes allow you to set the number of digits that MARK outputs for each parameter estimate, and to set the number of parameters and number of functions MARK will use in finding the maximum log likelihood.

Once you have checked the boxes you want, click OK. Mark will ask you if you want to use the identity matrix. Click OK. (We'll modify the design matrix in

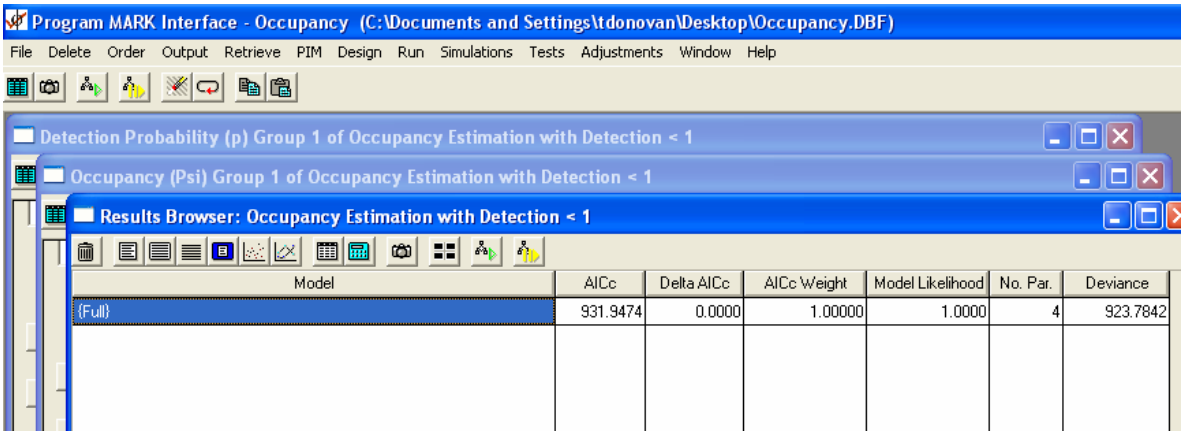
the covariate exercise). MARK will then run the analysis, and ask if you want to add the results to the results browser.



Click Yes.

THE RESULTS BROWSER IN MARK

The Results Browser will then appear with your model results.



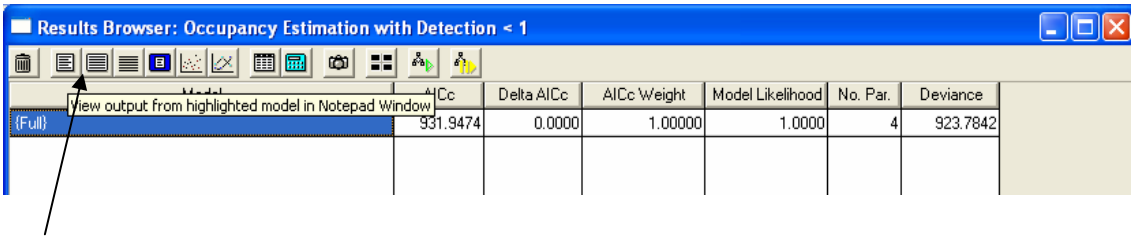
Let's study this output, and compare it to the spreadsheet results.

	E	F	G	H	I	J	K	L	M	N	O
15	OUTPUTS										
16	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-461.89	923.7841681	4	931.78	931.95	920.8247	2.9595	4	0.73987425	2.9792	0.5613

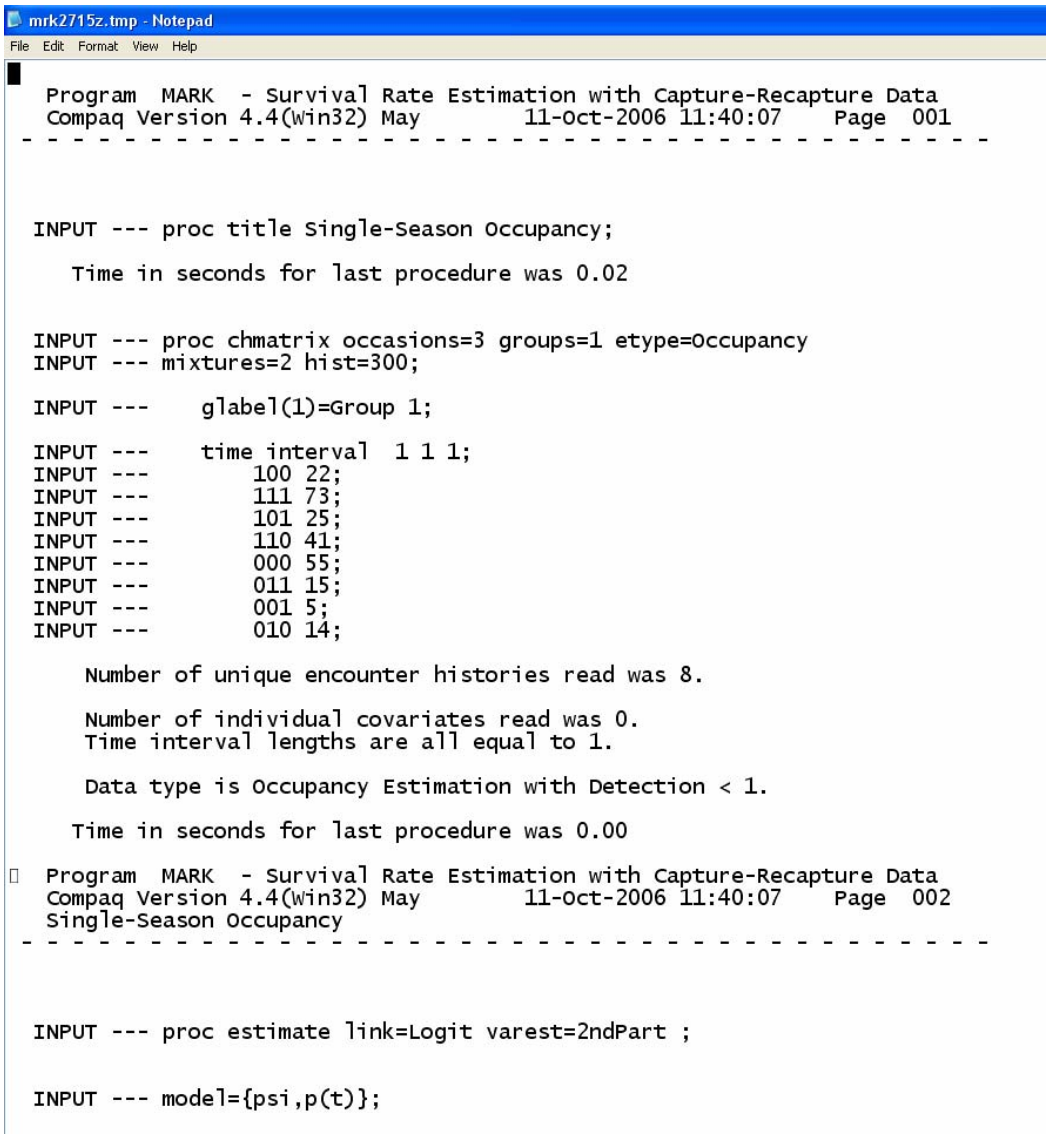
MARK reports the same AICc and K (No. Par.) as the spreadsheet (cells I17 and G17). You might remember that the spreadsheet computes Deviance as the difference between this model's -2Log$_e$L and the saturated model's -2Log$_e$L. However, as we'll soon see, the saturated model's -2Log$_e$L in MARK is 0, so the Deviance in MARK is the same thing as the model's -2Log$_e$L. Just keep in mind that for the occupancy models with no covariates, Deviance is the same thing as the -2Log$_e$L in MARK.

THE FULL OUTPUT

The Results Browser shows only some of the model output, and there are several ways to access more results, such as the beta and parameter estimates. The first way is to select the button to the right of the trash can, which in turn spawns a new NotePad window with <u>all</u> the results. The button to the right of this one will show only the beta estimates, and the button to the right of the beta button will show only the real parameter estimates. Most of these results can be exported into NotePad or into Excel.



Select the button to the right of the trash can to reveal the full output.



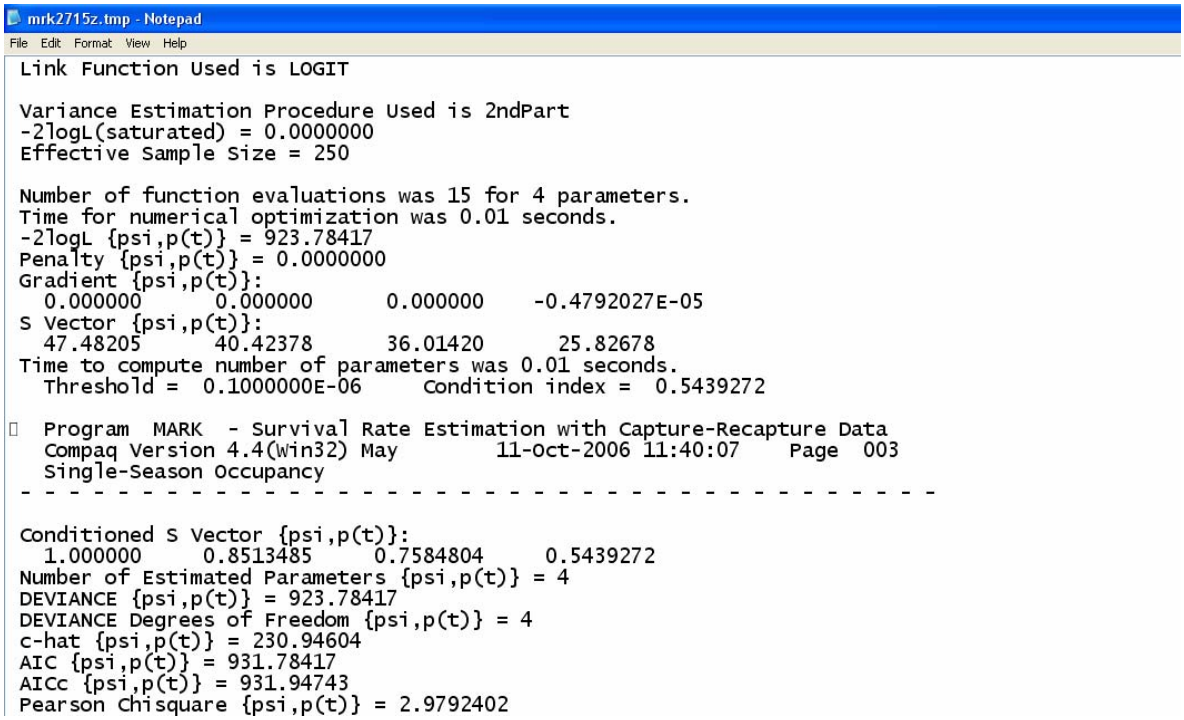
```
mrk2715z.tmp - Notepad
File  Edit  Format  View  Help

   Program  MARK  - Survival Rate Estimation with Capture-Recapture Data
   Compaq Version 4.4(Win32) May        11-Oct-2006 11:40:07    Page  001
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


  INPUT --- proc title Single-Season Occupancy;

     Time in seconds for last procedure was 0.02


  INPUT --- proc chmatrix occasions=3 groups=1 etype=Occupancy
  INPUT --- mixtures=2 hist=300;

  INPUT ---    glabel(1)=Group 1;

  INPUT ---    time interval  1 1 1;
  INPUT ---        100 22;
  INPUT ---        111 73;
  INPUT ---        101 25;
  INPUT ---        110 41;
  INPUT ---        000 55;
  INPUT ---        011 15;
  INPUT ---        001 5;
  INPUT ---        010 14;

      Number of unique encounter histories read was 8.

      Number of individual covariates read was 0.
      Time interval lengths are all equal to 1.

      Data type is Occupancy Estimation with Detection < 1.

     Time in seconds for last procedure was 0.00

   Program  MARK  - Survival Rate Estimation with Capture-Recapture Data
   Compaq Version 4.4(Win32) May        11-Oct-2006 11:40:07    Page  002
   Single-Season Occupancy
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


  INPUT --- proc estimate link=Logit varest=2ndPart ;


  INPUT --- model={psi,p(t)};
```

There's quite a bit of information in the full output, and we'll go through it step by step. The first sections describe a lot of the input that MARK read (the input file, the options you selected for running the analysis) and how it assigned labels to the output.

Scroll down to the section where MARK reports the link that was used in the analysis. It should be the logit link. Now look for the section of the output

that reports the -2Log$_e$L for the saturated model (0) – again it's computed differently than the spreadsheet.

Next, you should see the effective sample size = 250, which is the number of sites and is reported in cell F12 in the spreadsheet. Below this MARK spits out some information on how many functions it evaluated in its search for the multinomial maximum likelihood (in this case, 15). The Gradient, S Vector, and Threshold all deal with estimating the variance of the parameter estimates and how MARK counts the number of parameters that MARK actually estimated in the model.

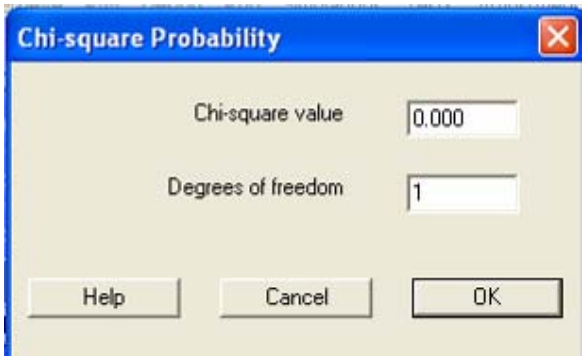


Below the dashed line, additional output is provided. The number of estimated parameters for this model was 4 (ψ, p_1, p_2, p_3, and p_4), which corresponds to cell G17 in the spreadsheet. The Deviance is again the difference between the -2Log$_e$L of this model and the saturated model, and since the saturated model's -

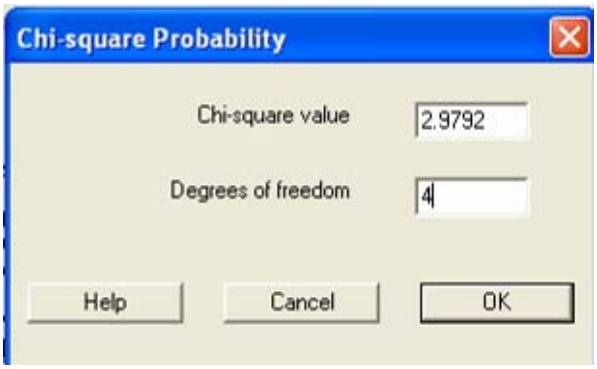
$2Log_eL = 0$ in MARK, Deviance is the same thing as the $-2Log_eL$ for the current model. The Deviance Degrees of Freedom is the same thing as the Model Degrees of Freedom (cell L17), and is 4 for this model. How was that computed? In the spreadsheet, we estimated it by counting the total kinds of histories and subtracting K. In this particular model, there are 8 kinds of encounter histories, thus the multinomial equation has 8 terms, or 8 parameters to be estimated (a probability for each history). In this occupancy model, we estimated those parameters through the estimation of ψ, p_1, p_2, p_3, and p_4 (4 occupancy parameters), leaving us with 8 – 4 = 4 parameters left. C-hat is estimated as 230.95, and is the model's Deviance/Model DF. The spreadsheet doesn't match the MARK output because Deviance is computed differently. The MARK result is simply the model's $-2Log_eL$/Model DF = 230.94604. The AIC and AICc scores from MARK match the spreadsheet.

THE PEARSON CHI-SQUARE OUTPUT

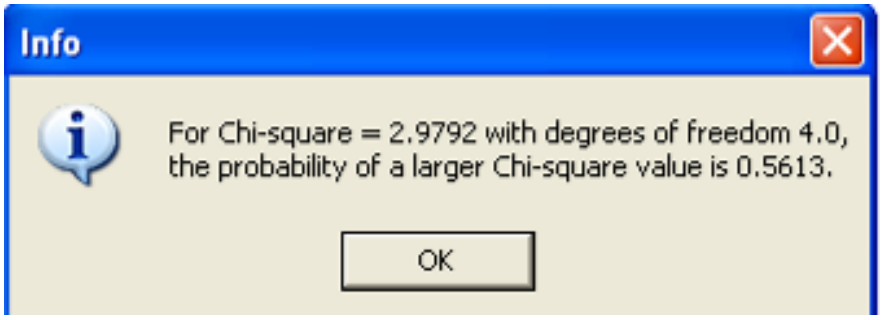
The last output in this section is the Pearson Chi-Square. MARK simply reports the Chi-Square test statistic value, and matches cell N17 in the spreadsheet. What about the p value? Well, you need to enter it and MARK will report the correct probability. Go to Tests | Chi-Square Prob, and a new dialogue box will appear:



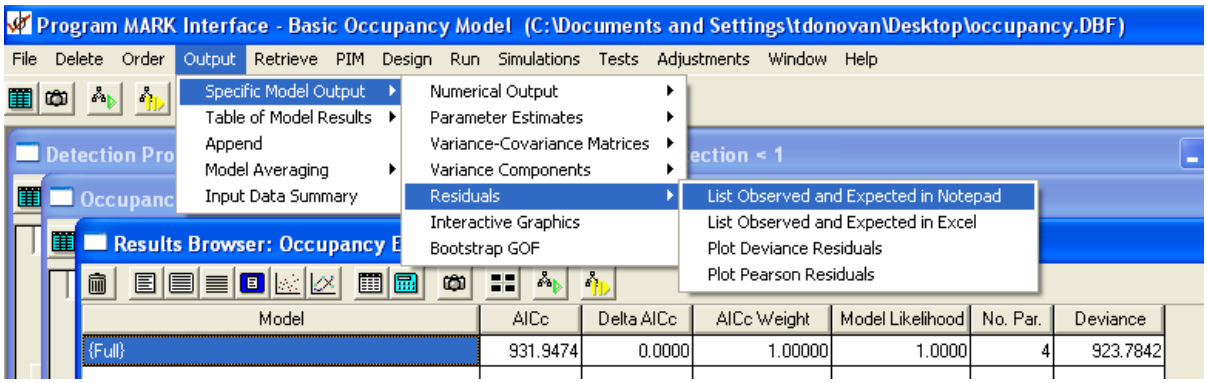
Here, copy the Chi-Square value from the output, and enter the Model DF as shown:



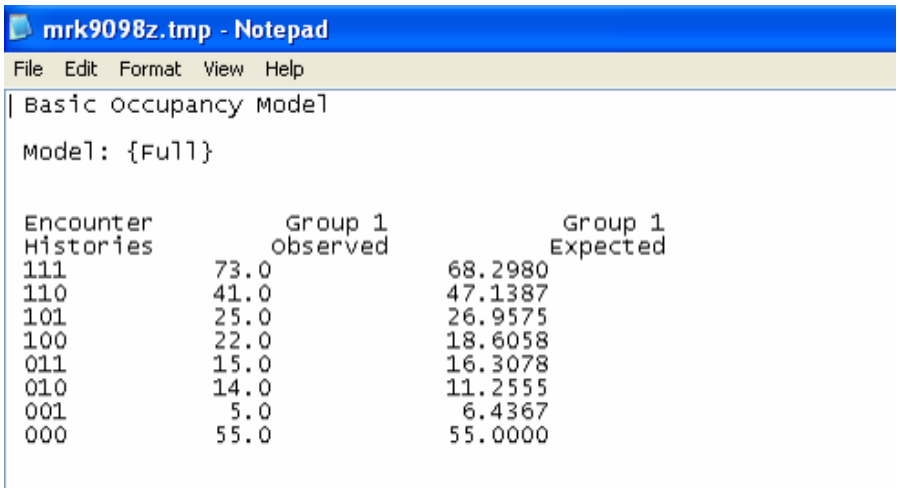
Press OK. MARK will open a new dialogue box with the result:



These results match N17 and O17 in the spreadsheet. You can determine how MARK computed the observed and expected values for the Pearson Chi-Square statistic by looking at the model residuals. Go to Output | Specific Model Output | Residuals | List Observed and Expected in Notepad (or Excel):



You should see the following output:



The encounter histories and their observed and expected values are provided, and these match the spreadsheet computations in cells F4:11 and L4:11.

	E	F	L	M
2				
3	History	Frequency	Expected	Chi-Square
4	100	22	18.61	0.62
5	111	73	68.30	0.32
6	101	25	26.96	0.14
7	110	41	47.14	0.80
8	000	55	55.00	0.00
9	011	15	16.31	0.10
10	001	5	6.44	0.32
11	010	14	11.26	0.67
12	# Sites =	250	250	2.97923663
13	# Histories =	8		

As a side note, outputting the residuals from MARK is a GREAT way to double-check your spreadsheet history probabilities for tricky histories. For example, given this model, the probability of getting a 000 history can be computed as its expected frequency divided by the total sites, or 55/250 = 0.22 (cell N8). If you develop your own spreadsheets, it's useful to know what answer you SHOULD get when you type in those long formulas for computing the probability of a history.

BETA AND REAL ESTIMATES

OK, back to the full output. Scroll down a bit more to find more important output, namely, the parameter estimates and standard errors:

mrk2715z.tmp - Notepad

File Edit Format View Help

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Conditioned S Vector {psi,p(t)}:
  1.000000     0.8513485     0.7584804     0.5439272
Number of Estimated Parameters {psi,p(t)} = 4
DEVIANCE {psi,p(t)} = 923.78417
DEVIANCE Degrees of Freedom {psi,p(t)} = 4
c-hat {psi,p(t)} = 230.94604
AIC {psi,p(t)} = 931.78417
AICc {psi,p(t)} = 931.94743
Pearson Chisquare {psi,p(t)} = 2.9792402


                        LOGIT Link Function Parameters of {psi,p(t)}
                                                              95% Confidence Interval
Parameter                    Beta         Standard Error      Lower           Upper
-------------------------  -------------  -------------  -------------  -------------
   1:p                     1.4322387       0.1897416      1.0603453       1.8041322
   2:p                     0.9296208       0.1626734      0.6107810       1.2484607
   3:p                     0.3707864       0.1469520      0.0827605       0.6588124
   4:Psi                   1.3724166       0.1680043      1.0431281       1.7017050


                        Real Function Parameters of {psi,p(t)}
                                                              95% Confidence Interval
Parameter                  Estimate       Standard Error      Lower           Upper
-------------------------  -------------  -------------  -------------  -------------
   1:p                     0.8072499       0.0295233      0.7427565       0.8586512
   2:p                     0.7169984       0.0330083      0.6481189       0.7770333
   3:p                     0.5916490       0.0355037      0.5206783       0.6589936
   4:Psi                   0.7977703       0.0271046      0.7394531       0.8457573

    Time in seconds for last procedure was 0.02

  Program  MARK  - Survival Rate Estimation with Capture-Recapture Data
   Compaq Version 4.4(Win32) May        11-Oct-2006 11:40:07    Page  004
   Single-Season Occupancy
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


 INPUT --- proc stop;

    Time in minutes for this job was 0.00

         E X E C U T I O N   S U C C E S S F U L
```

Now you can see how the PIM indexing is reflected in the output. In the PIM chart, we identified p_1 as Parameter 1, p_2 as Parameter 2, and so on. The parameters are identified as 1:p in the MARK output, indicating that parameter 1 is p (the first p). The betas, standard errors of the betas, and 95% confidence limits are provided on the top panel, and the estimates, standard errors, and 95% confidence limits are provided for the real parameters in the last section. Remember that MARK "worked" on the beta estimates to find the maximum log likelihood, and then back-transformed the betas to the real estimates with a logit link. Here are the spreadsheet results:

	G	H	I	J
3	Parameter	Estimate?	Betas	MLE
4	p1	1	1.432238	0.80725
5	p2	1	0.92962	0.71700
6	p3	1	0.370786	0.59165
7	ψ	1	1.372416	0.79777

You can see that the spreadsheet matches the MARK, although MARK reports the standard errors while the spreadsheet (currently) does not. In this section of the output, you'll want to scrutinize the standard errors very carefully because here you'll get an indication of whether MARK had any trouble estimating a parameter. If you see standard errors or confidence intervals that are completely unreasonable, it indicates some sort of estimation problem.

MODEL PSI(.)P(.)

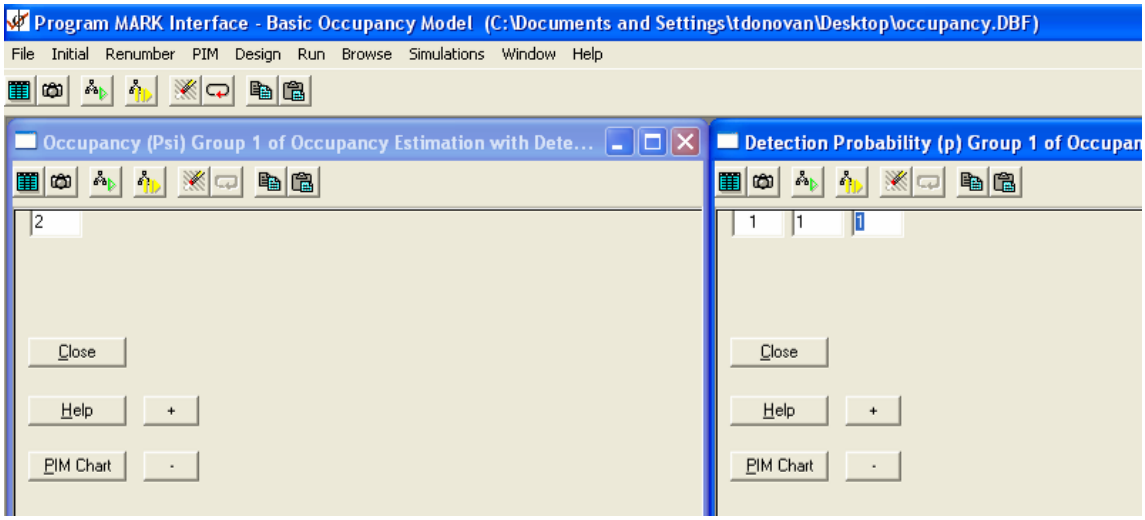
Now let's run another model, one where p is constant. In the spreadsheet, we set this model up as follows:

	G	H	I
3	Parameter	Estimate?	Betas
4	p1	1	
5	p2	0	=I4
6	p3	0	=I4
7	ψ	1	

and let Solver find beta values in cells I4 and I7 that maximized the multinomial log likelihood. In this model, we are interested in estimating two parameters only, and will force $p_1 = p_2 = p_3$. Let's see how this would be set up in MARK.

Go to PIM | Open Parameter Index Matrix. The default (the model we just ran) is the full model, where we estimated all four parameters. In this model,

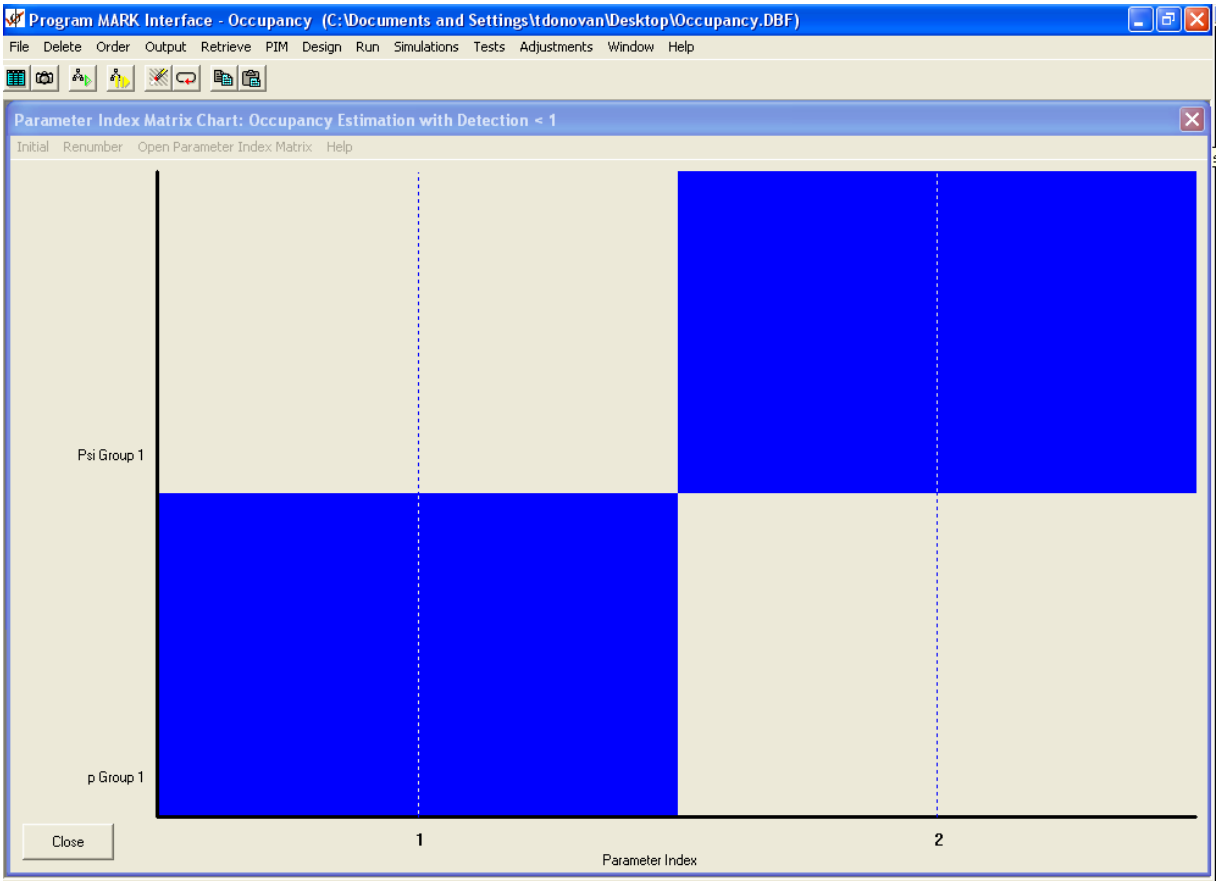
however, we need to estimate just two, and we need to tell MARK that the p's will be equal. You do that by forcing all three p estimates to have the same Parameter Index. Below, we entered a 1 for p_1, p_2, and p_3 in the PIM, indicating that all three p's are equal, and that they will be referenced as Parameter Index 1. We entered a 2 in the occupancy window to tell MARK that we will estimate a separate beta for ψ, and that it will be referenced as Parameter Index 2.



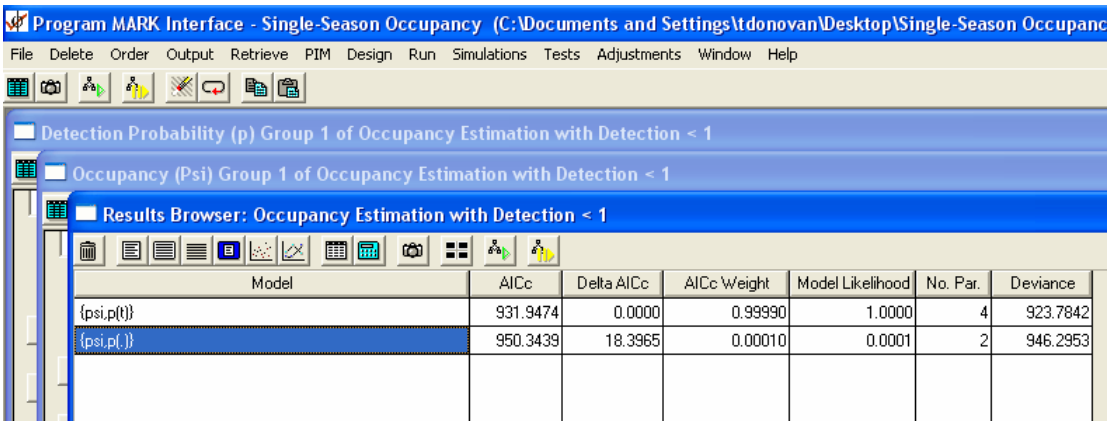
You can use the + and – buttons to help enter these more quickly, and also Gary White has built in many options in MARK so that when right-click on the PIM window, several short-cuts are available. (You'll have to play around with these on your own).

The other way to parameterize model Psi(.)P(.) is to use the PIM Chart. Go to PIM | Open Parameter Index Chart. Right click somewhere on the p estimates blue box, and select "Constant" – indicating that you are forcing $p_1 = p_2 = p_3$. Right click again and select "Renumber with Overlap"…you should see the

following PIM chart. Note that there are now only two estimates: p and psi, and p will be labeled "parameter 1" and psi will be labeled "parameter 2."



Then go to Run | Current Model, and give this new model the name Psi(.)P(.) – which indicates that there is now a single estimate for psi and p. Click OK to Run and append the results to the database.



MODEL PSI(.)P(.) OUTPUT

This new model has a log likelihood of -473.16, a $-2\text{Log}_e\text{L}$ of 946.2953, and an AICc score of 950.3439. K = 2 because there are only two parameters estimated in this model. Retrieve this model by clicking on the button with the four black squares on the toolbar. Then run your spreadsheet Solver with the constraints that cells I5:I6 = I4. You should see that the spreadsheet output matches the output from MARK.

	E	F	G	H	I	J	K	L	M	N	O
15	OUTPUTS										
16	Log_eL	$-2\text{Log}_e\text{L}$	K	AIC	AICc	$-2\text{Log}_e\text{L}$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-473.15	946.2952919	2	950.30	950.34	920.8247	25.4706	6	4.245103459	24.4334	0.0004

Take some time now and go through the MARK output, such as the K, AIC, Model DF, c-hat, Chi-Square, and P, and compare MARK's output with the spreadsheet. This is an important step because it forces you to recognize how MARK is generating the output that it is.

Now let's look at the beta estimates and parameter estimates:

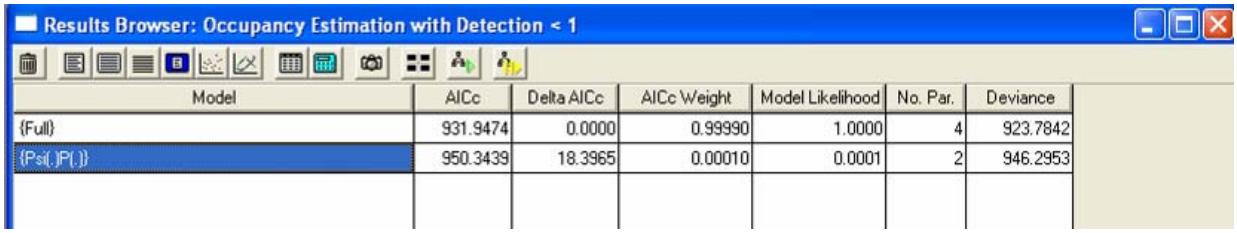
```
                      LOGIT Link Function Parameters of {psi,p(.)}
                                                                      95% Confidence Interval
 Parameter                          Beta         Standard Error      Lower            Upper
 -------------------------  --------------  --------------  --------------  --------------

   Program  MARK  - Survival Rate Estimation with Capture-Recapture Data
   Compaq Version 4.4(Win32) May        11-Oct-2006 12:00:30    Page  003
   Single-Season Occupancy
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

    1:p                         0.8584824       0.0992453       0.6639615       1.0530032
    2:Psi                       1.3933561       0.1712806       1.0576462       1.7290660


                      Real Function Parameters of {psi,p(.)}
                                                                      95% Confidence Interval
 Parameter                          Estimate     Standard Error      Lower            Upper
 -------------------------  --------------  --------------  --------------  --------------
    1:p                         0.7023435       0.0207479       0.6601497       0.7413512
    2:Psi                       0.8011275       0.0272888       0.7422405       0.8492929

     Time in seconds for last procedure was 0.02


  INPUT --- proc stop;

     Time in minutes for this job was 0.00


          E X E C U T I O N   S U C C E S S F U L
```

	E	F	G	H	I	J	K	L	M	N	O
1	Single-Species, Single-Season Model										
2							Probability			Saturated Model	
3	History	Frequency	Parameter	Estimate?	Betas	MLE	of History	Expected	Chi-Square	Probability	Ln(Prob)
4	100	22	p1	1	0.858482	0.70234	0.050	12.46	7.30	0.088	-2.43041846
5	111	73	p2	0	0.858482	0.70234	0.278	69.39	0.19	0.292	-1.23100148
6	101	25	p3	0	0.858482	0.70234	0.118	29.41	0.66	0.1	-2.30258509
7	110	41	ψ	1	1.393356	0.80113	0.118	29.41	4.57	0.164	-1.80788885
8	000	55					0.220	55.00	0.00	0.22	-1.51412773
9	011	15					0.118	29.41	7.06	0.06	-2.81341072
10	001	5					0.050	12.46	4.47	0.02	-3.91202301
11	010	14					0.050	12.46	0.19	0.056	-2.88240359
12	# Sites =	250					1	250	24.43335294	Log L (sat)	-460.41234
13	# Histories =	8								-2 Log L (sat)	920.82467
14											
15	OUTPUTS										
16	Log$_e$L	-2Log$_e$L	K	AIC	AICc	-2Log$_e$L Sat	Deviance	Model DF	C-hat	Chi-Square	P value
17	-473.15	946.2952919	2	950.30	950.34	920.8247	25.4706	6	4.245103459	24.4334	0.0004

You can see that Solver found the same estimates as MARK. The goal of the spreadsheet exercise is simply to reinforce the multinomial log likelihood, and to show how different output is calculated.

One thing that we did not do on the spreadsheet was to obtain model weights (you can of course add that if you want to!). One nice thing about MARK is that

it keeps track of these results for you, and you can easily export the results. The Browser allows you to compare the two models we ran easily.



Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance
{Full}	931.9474	0.0000	0.99990	1.0000	4	923.7842
{Psi(.)P(.)}	950.3439	18.3965	0.00010	0.0001	2	946.2953

Notice that the full model has the lowest AICc score, and that the Psi(.)P(.) model has a delta AICc score of 18.3965. The model weights are computed from these delta scores, and the full model has a weight of 0.9999, and the p(.) model has a weight of 0.0001. This is good because, as you might recall, the data were simulated with $p_1 = 0.8$, $p_2 = 0.7$, and $p_3 = 0.6$! (See columns Q:W on the spreadsheet for how to simulate new data). We would conclude that the full model has a 0.9999 probability of being the best K-L model in this model set. We'll cover model selection and model averaging in much more detail in the covariate exercise.

That's all for this exercise! You've run an occupancy model in MARK, and hopefully understand exactly how MARK is estimating basic occupancy parameters. In the next exercise, we'll see how to force the occupancy parameters to be functions of covariates.