EXERCISE 12: SINGLE-SPECIES, SINGLE-SEASON OCCUPANCY MODELS WITH IDENTIFICATION ERRORS

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 WITH IDENTIFICATION ERRORS

OBJECTIVES:

- To learn and understand the single-season occupancy model that assesses false positive detections, and how it fits into a multinomial maximum likelihood analysis.
- To use Solver to find the maximum likelihood estimates for the probability of false positive detection, and the probability of detection and site occupancy for each group.
- To assess the -2Log$_e$L of the saturated model.
- To introduce concepts of model fit.
- To learn how to simulate single-season occupancy data with false positive identifications.

BASIC INFORMATION

If you've been completing the exercises in this book in order, you've learned a great deal about the single-season occupancy modeling, and some interesting variations of the basic model. In this exercise, we describe occupancy models in which false, positive identifications are included in the dataset. This model was developed by Andy Royle and Bill Link, and is described in the paper: Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835-841. Click on the worksheet labeled "Errors" and we'll get started.

BACKGROUND

Hopefully by now you have a solid understanding that the general occupancy model handles the fact that an encounter history full of zeroes (e.g., "0000") can indicate either that the species of interest was absent from a site, or that species was present but simply undetected. You might recall that the encounter history probability for such records handles this by adding the two possible outcomes together: $\psi * (1-p_1) * (1-p_2) * (1-p_3) + (1-\psi)$. In this way, the generalized site occupancy model allows for false negative observations; i.e., observations recorded as 0 but should have been recorded as 1. The first term deals directly with false negative errors, while the second term handles sites that are truly unoccupied (no error; a 0 is correctly recorded).

There is another kind of error, however, which might be made by field observers, and those are false positive errors; i.e., observations recorded as 1 but should have been recorded as 0. These kinds of errors almost certainly occur. Andy Royle and Bill Link point out that "even low false positive error rates can induce extreme bias in estimates of site occupancy when they are not accounted for." Here's a quick example (based on personal experience). When avian ecologists conduct surveys of breeding birds, the vast majority of detections are obtained when the observer hears the song or calls issued by individuals in the early morning hours. Some species are notoriously hard to differentiate. For instance, red-eyed vireos and blue-headed vireos are very similar in song except for differences in the singing rate and "sweetness." In such cases, it is fairly easy to misidentify one species as the other.

You might recall that the single-season occupancy model assumes that these kinds of errors are not made, but false positives can be more common than anyone would like to admit. Nonetheless, they do occur and need to be handled.

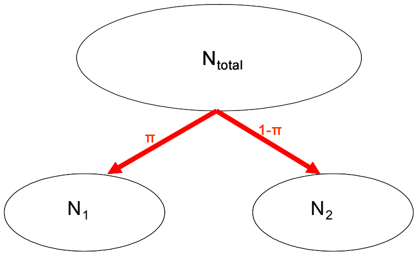
So, how can the single-season occupancy model be extended to account for false positive detections? Well, believe it or not, the model that Andy Royle and Bill Link used as a starting point was the mixture model. So, let's quickly review the mixture model, and then we will extend the concepts from the basic mixture model to include false errors.

REVIEW OF SINGLE SEASON MIXTURE MODELS

The idea behind mixture models, also called heterogeneity models, is that sites in the study area are unique in some way, such that there is heterogeneity among sites in terms of detection and occupancy probability. Mixture models are used when investigators either don't record or don't have access to site- or survey-level covariates. Thus, in mixture models, all of the differences among sites are unobservable or unknown.

Unobservable heterogeneity refers to situations when the factors causing differences in either occupancy probability or detection probability cannot be readily identified. This could simply mean we have absolutely no clue what might cause differences, but are willing to accept that there might be differences that we cannot measure. For instance, if food resources are a critical predictor of occupancy but cannot be measured readily across sites, it might impose heterogeneity among the study sites, where some sites are rich in food resources and others are poor, even though food was not measured directly.

So, how does one model unobservable heterogeneity? Well, the basic idea is that the study sites can be divided into multiple groups, and each group (not each site) has unique detection probabilities and a unique probability of occupancy. The number of groups can be either a discrete number (e.g., 2 groups, 3 groups, etc.) or an infinite number.



In exercise 6, we focused on a heterogeneity model in which the group number is discrete (n = 2), and heterogeneity was modeled for detection probability only. So, for the two-point mixture model we divide the study sites into two groups. The population of study sites (N_{total}) is divided into group 1 and group 2 such that $N_{total} = N_1 + N_2$, where N_1 is the total number of sites in group 1, and N_2 is the total number of sites in group 2. The first trick is to figure out the proportion of sites that belong to group 1 and the proportion of sites that belong to group 2. We let π represent the proportion of sites in group 1, so by definition $N_1 = \pi N_{total}$. Because there are only two groups being considered, if we know N_1 then we can derive N_2 as $N_2 = (1 - \pi) N_{total}$ because a site must be in group 2 if it wasn't in group 1. After we estimate the proportion of the population in each group (π and 1-π), the next step is to estimate the detection and occupancy probabilities for each group separately.

Group 1: proportion of sites in group 1 = πN_{total}, and we estimate detection parameters p_1, p_2, p_3, p_4 and the occupancy parameter ψ specific to group 1.

Group 2: proportion of sites in group 1 = $(1-\pi)N_{total}$, and we estimate detection parameters, p_1, p_2, p_3, p_4 and the occupancy parameter ψ specific to group 2. (Note: recall that ψ_2 must be equal to ψ_1; occupancy cannot be group-specific or the model is unidentifiable.)

You might recall that the encounter histories are then written out for each group separately. For example, suppose we conduct an occupancy survey in which sites are surveyed 4 times. In the single-season, single-species, 1 group model, the probability of realizing a 1111 history is:
Probability of 1111 = $\psi\ p_1\, p_2\, p_3\ p_4$.

OK, now let's extend this to a two-group mixture:
Probability of 1111 = $\pi\ \psi_1\, p_{1,1}\, p_{2,1}\, p_{3,1}\ p_{4,1} + (1-\pi)\ \psi_2\, p_{1,2}\, p_{2,2}\, p_{3,2}\, p_{4,2}$

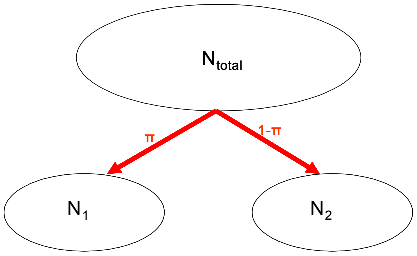
where the first subscript indicates the survey period, followed by a comma, and the second subscript refers to the group membership. In other words, the probability of obtaining a 1111 history is the sum of two probability terms, one for group 1 and the second for group 2. The probability of getting a 1111 history for group 1 is $\pi\ \psi_1\, p_{1,1}\, p_{2,1}\, p_{3,1}\ p_{4,1}$. The site must first be part of group 1 (π), the site must be occupied (ψ_1), the species must be detected on the first survey ($p_{1,1}$), the species must be detected on the second survey ($p_{2,1}$), the species must be detected on the third survey ($p_{3,1}$), and the species must be detected on the fourth survey ($p_{4,1}$). All of these terms are multiplied together because all of them must occur to generate a 1111 history for a site in group 1. The probability of getting a 1111 history for group 2 is $(1-\pi)\ \psi_2\ p_{1,2}\ p_{2,2}\ p_{3,2}\ p_{4,2}$. The site must first be part of group 2 $(1-\pi)$, the site must be occupied (ψ_2), the species must be detected on the first survey ($p_{1,2}$), the species must be detected on the second survey ($p_{2,2}$), the species must be detected on the third survey ($p_{3,2}$), and the species must be detected on the fourth survey ($p_{4,2}$). All of these terms are multiplied together because all of them must occur to generate a 1111 history for a site in group 2. The two terms are added together because a site can be in either group 1 (in which case the parameters apply to group 1) OR it can be in group 2 (in which case the occupancy parameters apply to group 2).

OK, let's go through just one more, 0000. In a one-group, single season model, the probability is $\psi\ (1-p_1)\ (1-p_2)\ (1-p_3)\ (1-p_4)+(1-\psi)$. The site could have been occupied but missed on all four surveys (a false negative error), OR it could have been unoccupied. Extending this to two group mixture model, the probability of a 0000 history would have four terms, the first two apply to sites in group 1, while the last two apply to sites in group 2: $\pi\ \psi\ (1-p_{1,1})\ (1-p_{2,1})\ (1-p_{3,1})\ (1-p_{4,1}) + \pi\ (1-\psi) + (1-\pi)\ \psi\ (1-p_{1,2})\ (1-p_{2,2})\ (1-p_{3,2})\ (1-p_{4,2}) + (1-\pi)\ (1-\psi)$.

Finally, it's a good idea to recall that mixture models are data hungry and care must be taken to not overparameterize the model. Additionally, there is the additional, sticky issue that the likelihood surface may have local optima…we discussed both of the issues in detail in exercise 6.

EXTENDING THE MODEL TO INCLUDE FALSE-POSITIVES

As we mentioned previously, the two-point mixture model is the basis for the occupancy model in which false-positives and false negatives can be determined. Again, the sites are divided into two groups. But this time, the dividing factor is whether the sites are truly occupied or truly vacant. In the diagram to the right, this means that N_1 sites are truly occupied, and N_2 sites are truly empty.



Now we write out encounter history probabilities for each group, starting with group 1 (sites are truly occupied). We'll stick with our study in which sites were surveyed four times. The probability of realizing a 1111 history is: ψ*p_1*p_2*p_3*p_4. The probability of realizing a 1010 history is: ψp_1*(1-p_2)*p_3*(1-p_4). In other words, for sites that are truly occupied, the encounter history probabilities are exactly like those in the single-season occupancy model, with one exception. The exception is the history which contains all zeroes. In this case, the probability of realizing a 0000 history is ψ*(1-p_1)*(1-p_2)*(1-p_3)*(1-p_4). Because group 1 consists of only sites that are truly occupied, the term (1-ψ) is dropped.

OK, now let's look at group 2, which consists of sites that are truly unoccupied. If sites are truly empty, then any histories with a "1" in them clearly have errors. For history 0001, a false positive error was made on survey 4. For history 0101, a false positive error was made on surveys 2 and 4. For history 1111, four false positives were recorded. Holy crow!

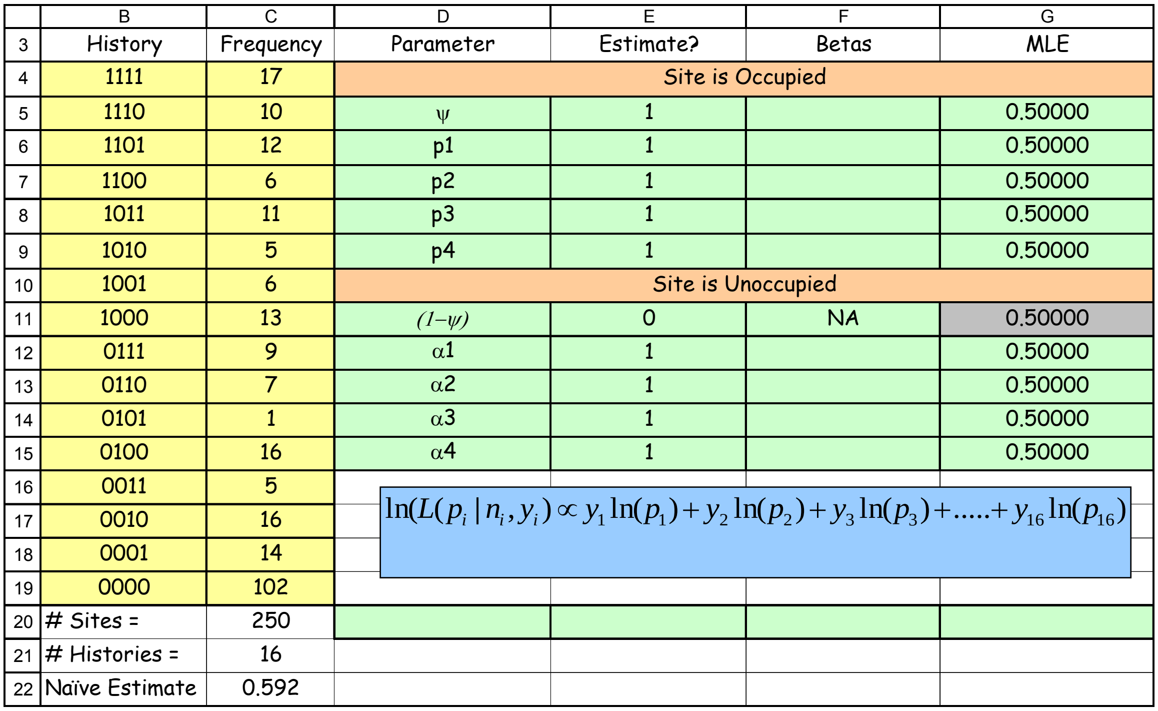
To write out the encounter history probabilities for group 2, we need a new parameter, called alpha (or α). Alpha is the probability of detecting a species of interest when it is in fact absent. (You can think of alpha as the opposite of p, which is the probability of detecting a species given it is present). Note: Royle and Link refer to this parameter p_{10}, but we like to call it alpha to avoid confusion with detection probability p). Alpha can be survey specific, so in our four survey example, we can estimate up to four new parameters for the error model: α_1, α_2, α_3, and α_4.

OK, let's look at some histories for group 2 (sites that are truly unoccupied). In this group, the probability of realizing a 1111 history is $(1-\psi)*\alpha_1*\alpha_2*\alpha_3*\alpha_4$. The site was not occupied $(1-\psi)$, and a mistake was made on survey 1 (α_1), another false positive was made on survey 2 (α_2), another false positive was made on survey 3 (α_3), and another false positive was made on survey 4 (α_4). Given that the site is truly absent, the probability of realizing a 1010 history is $(1-\psi)*\alpha_1*(1-\alpha_2)*\alpha_3*(1-\alpha_4)$. The site was unoccupied $(1-\psi)$, a false positive was recorded in survey 1 (α_1), a correct 0 was recorded on survey 2 $(1-\alpha_2)$, a false positive was recorded in survey 3 (α_3), and a correct 0 was recorded on survey 4 $(1-\alpha_4)$. As a final example, given that the site is truly unoccupied, the probability of realizing a 0000 history is $(1-\psi)*(1-\alpha_1)*(1-\alpha_2)*(1-\alpha_3)*(1-\alpha_4)$. Hopefully this makes sense to you.

As with the two-point mixture model, we add the encounter history probabilities for both groups. For instance, the probability of observing a 1111 history is $\psi*p_1*p_2*p_3*p_4 + (1-\psi)*\alpha_1*\alpha_2*\alpha_3*\alpha_{4,}$ where the first term represents the probability for group 1 and the second term represents the probability for group 2. The probability of observing a 1010 history is $\psi p_1*(1-p_2)*p_3*(1-p_4) + (1-\psi)*\alpha_1*(1-\alpha_2)*\alpha_3*(1-\alpha_4)$. The probability of observing a 0000 history is $\psi*(1-p_1)*(1-p_2)*(1-p_3)*(1-p_4) + (1-\psi)*(1-\alpha_1)*(1-\alpha_2)*(1-\alpha_3)*(1-\alpha_4)$. Given the full set of possible histories and the frequencies, the goal of the analysis is to use maximum likelihood procedures to estimate ψ, p_1, p_2, p_3, p_4, α_1, α_2, α_3, and α_4. That's 9 parameters to estimate, and because a four survey study results in 16 possible encounter histories, you don't have to worry about overparameterization. (You can estimate up to 15 parameters validly).

ERROR MODEL SPREADSHEET INPUTS

OK, with that background, let's get oriented to the spreadsheet. In this example, the investigator surveys 250 study sites, with each site being surveyed 4 times. The encounter histories are recorded in cells B4:B19, and the frequency of each history is recorded in cells C4:C19. The total number of sites is given in cell C20, and the number of unique histories is given in cell C21 (which you might remember indicates the number of terms in our multinomial likelihood function). To avoid over-parameterization, you can only run models with 15 or fewer parameters. The naïve estimate for occupancy (occupancy unadjusted for detection probability) is computed in cell C22 as the total number of sites which had one or more detections divided by the total number of sites. In this case the estimate is around 60%. Keep that in mind as we move through the exercise.



	B	C	D	E	F	G
3	History	Frequency	Parameter	Estimate?	Betas	MLE
4	1111	17	Site is Occupied			
5	1110	10	ψ	1		0.50000
6	1101	12	p1	1		0.50000
7	1100	6	p2	1		0.50000
8	1011	11	p3	1		0.50000
9	1010	5	p4	1		0.50000
10	1001	6	Site is Unoccupied			
11	1000	13	$(1-\psi)$	0	NA	0.50000
12	0111	9	α1	1		0.50000
13	0110	7	α2	1		0.50000
14	0101	1	α3	1		0.50000
15	0100	16	α4	1		0.50000
16	0011	5	$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + ... + y_{16} \ln(p_{16})$			
17	0010	16				
18	0001	14				
19	0000	102				
20	# Sites =	250				
21	# Histories =	16				
22	Naïve Estimate	0.592				

OK, now let's look at the parameters. Notice the spreadsheet is divided into two sections. In the first section (cells D4:G9), we list the parameters associated with group 1, consisting of sites that are truly occupied (ψ, p_1, p_2, p_3, p_4). The second section of the spreadsheet (cells D10:G10), lists the parameters (1-ψ, α_1, α_2, α_3, α_4) for group 2, which consists of sites that are truly unoccupied. As with other spreadsheet exercises, you enter a 1 when a parameter is being uniquely estimated, or enter a 0 if the parameter is being forced to be equal to some other parameter. Note that we won't estimate (1-ψ) because we can derive it from the estimate of ψ, so a 0 is entered in cell E11 and the MLE in cell G11 is computed as 1-G5.

MIXTURE LINKS

The betas for each parameter are listed in cells F5:F9, F12:F15, and the parameter estimates that correspond to each beta are computed in cells G5:G9, G12:G15 through a logit link. Click on cell G5 and you'll see the logit link transformation: =EXP(F5)/(1+EXP(F5)). Remember the logit link constrains the MLE's to be between 0 and 1, which is what we want for ψ, and the p_i's, and the α_i's. Because this exercise doesn't include covariates, we also could have used the sin link, but we stuck with the logit in case we decide to add covariates some day.

Keep in mind that we don't know what the beta values are.....we are going to let Solver find the betas that maximize the multinomial log likelihood function (see below).

SPREADSHEET HISTORY PROBABILITIES

OK! Now we are ready to compute the probability of realizing each history. Let's start with the first history listed, 1111, in cell B4.

The probability of realizing a 1111 history is estimated for each group separately. If sites are truly occupied, the probability of realizing a 1111 history is $\psi * p_1 * p_2 * p_3 * p_4$. This equation is entered in cell H4: =G5*G6*G7*G8*G9. If sites are truly unoccupied, the probability of realizing a 1111 history is $(1- \psi) * \alpha_1 * \alpha_2 * \alpha_3 * \alpha_4$. This equation is entered in cell I4: =G11*G12*G13*G14*G15. Across both groups, the probability of realizing a 1111 history is the sum of the two mixing probabilities, given in cell J4 (=H4+I4). The natural log of the combined history probabilities is computed in cell K4. And so it goes for the remaining histories.

Make sense? Spend time now clicking on the formula for each history and for group. In our experience, if students understand how the encounter histories are calculated, the rest is a piece of cake.

Notice that the sum of cells J4:J19 must equal 1 (cell J20): there are 16 possible histories, and each history has a probability of being realized, but the sum of the probabilities must be 1.00.

THE ERROR MODEL MULTINOMIAL LOG LIKELIHOOD

The goal of the analysis, as you might have guessed, is to find the combination of betas that maximizes the multinomial log likelihood function. Remember, by changing the betas, we change the parameter estimates linked to each beta, which changes the probability of each encounter history, which changes the Log$_e$L.

Betas → MLEs → Encounter Histories → Log$_e$L

All that's left is to compute the log likelihood, given the frequencies of each history and the history's probability. The multinomial log likelihood formula that we've been using is in the blue box below.

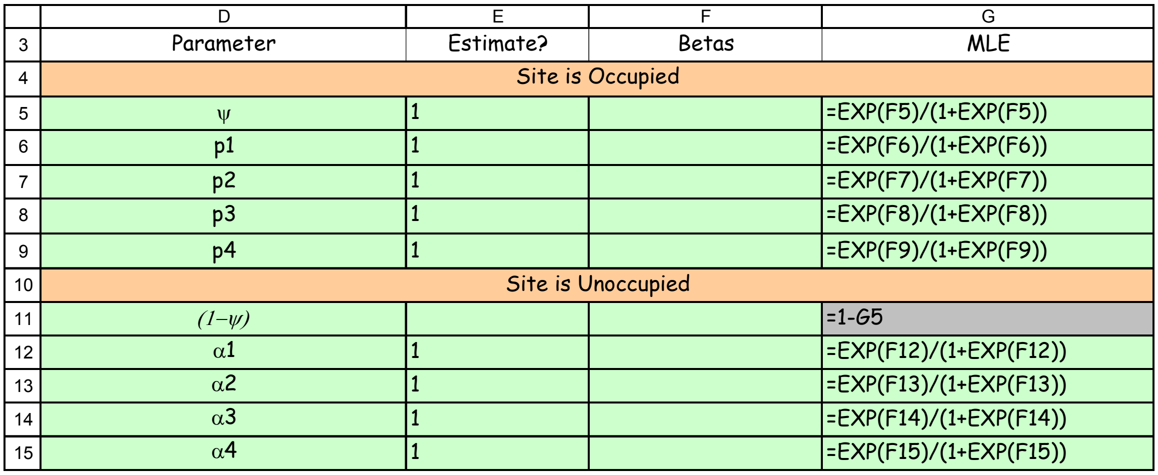
$$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + \ldots.. + y_{16} \ln(p_{16})$$

There are 16 terms in this function, one for each of the encounter histories. The y_i in the blue box are the frequencies of each kind of history and the p_i

in the blue box equation above are the history probabilities. The Log$_e$L is computed in cell B26 with the equation =SUMPRODUCT(C4:C19,K4:K19), which corresponds to the general formula in the blue box. Now all we have to do is maximize this value to find the MLE's for our dataset.

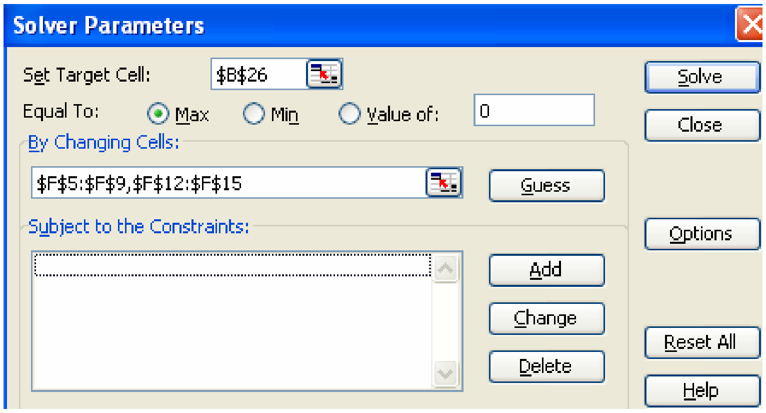
MAXIMIZING THE LOG LIKELIHOOD

Before we run our first model, we need to make sure that $(1-\psi_1)$ is correctly entered in cell G11, so set your spreadsheet up as follows:



	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1		=EXP(F5)/(1+EXP(F5))
6	p1	1		=EXP(F6)/(1+EXP(F6))
7	p2	1		=EXP(F7)/(1+EXP(F7))
8	p3	1		=EXP(F8)/(1+EXP(F8))
9	p4	1		=EXP(F9)/(1+EXP(F9))
10	Site is Unoccupied			
11	$(1-\psi)$			=1-G5
12	α1	1		=EXP(F12)/(1+EXP(F12))
13	α2	1		=EXP(F13)/(1+EXP(F13))
14	α3	1		=EXP(F14)/(1+EXP(F14))
15	α4	1		=EXP(F15)/(1+EXP(F15))

Make sure that the beta cells are cleared out. OK, now we're ready to run this model. We can call this model "ψ, p(t)a(t)" to indicate that we're estimating ψ, plus p_1, p_2, p_3, and p_4, and α_1, α_2, α_3, α_4. You know the drill. Open Solver, and set cell B26 to a maximum by changing cells F5:F9, F12:F15.



Solver Parameters

Set Target Cell: $B$26

Equal To: ● Max ○ Min ○ Value of: 0

By Changing Cells: $F$5:$F$9,$F$12:$F$15

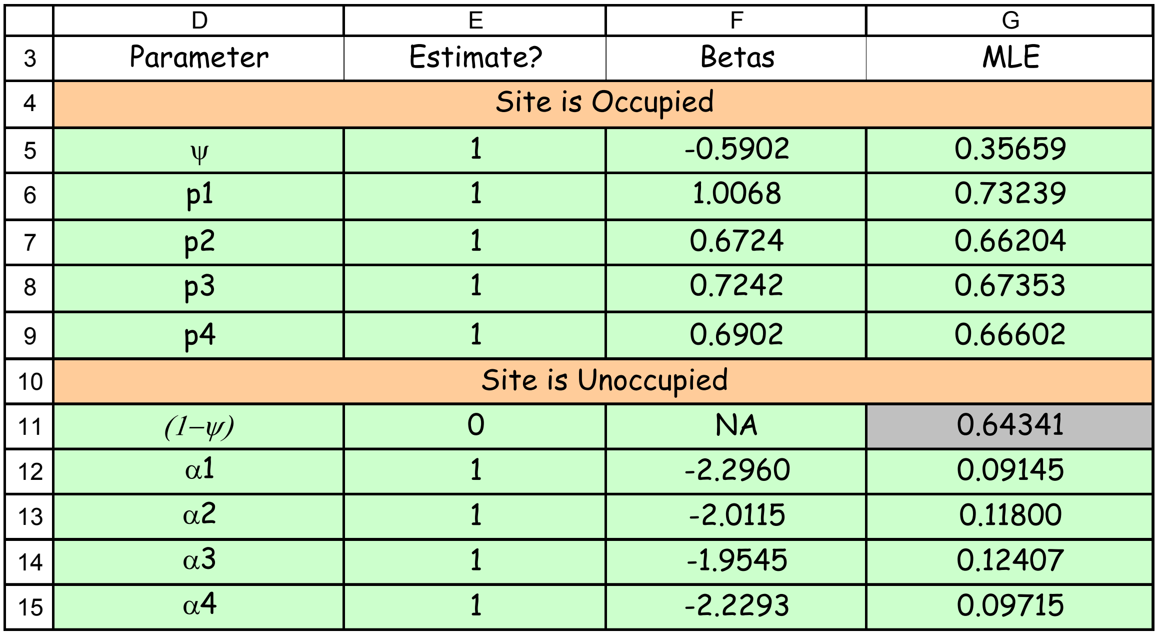
Subject to the Constraints:

Press Solve and Solver will work through the various combinations of betas until it finds the maximum.

ERROR MODEL OUTPUT

Before we study the output, it's important to note that the sometimes the parameters for the two groups get switched, meaning that group 1 is really group 2 and vica versa. If your results don't match the results shown on the next few pages, try "seeding" the betas with random numbers (e.g., enter =RAND() in a beta cell), and then try running Solver again…..with random starting betas, your groups might switch back again.

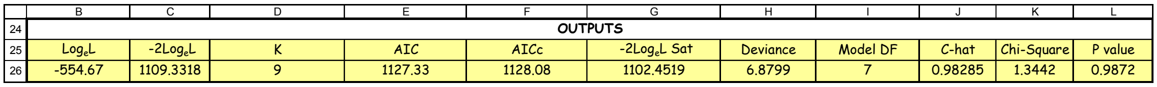
First, let's take a look at the parameter estimates found by Solver:



	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1	-0.5902	0.35659
6	p1	1	1.0068	0.73239
7	p2	1	0.6724	0.66204
8	p3	1	0.7242	0.67353
9	p4	1	0.6902	0.66602
10	Site is Unoccupied			
11	$(1-\psi)$	0	NA	0.64341
12	$\alpha 1$	1	-2.2960	0.09145
13	$\alpha 2$	1	-2.0115	0.11800
14	$\alpha 3$	1	-1.9545	0.12407
15	$\alpha 4$	1	-2.2293	0.09715

The proportion of sites that were truly occupied (group 1) is 0.35659 (cell G4). By subtraction, the proportion of sites that were truly unoccupied (group 2) is (1- 0. 35659) = 0.64341. These estimates would almost certainly be different if we assumed no false-positive errors. For occupied sites, p_1 = 0.73239 (cell G6), p_2 = 0.66204 (cell G7), p_3 = 0.67353 (cell G8), and p_4 = 0.66602. For sites that were truly unoccupied, the probability of making a false positive observation is fairly high across all four surveys: α_1 = 0.09145 (cell G12), α_2 = 0.1180 (cell G13), α_3 = 0.12407 (cell G14), and α_4 = 0.09715 (cell G15). Whether this model is supported by the data remains to be seen.

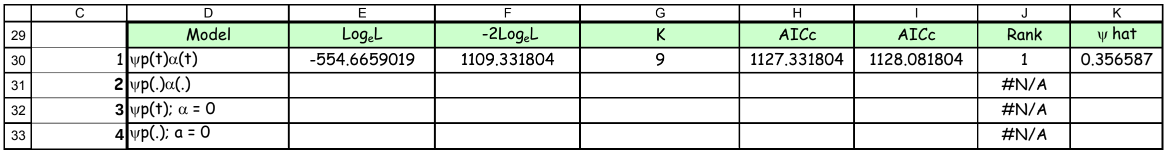
Now let’s look at the remaining output given in cells B25:L26.



	B	C	D	E	F	G	H	I	J	K	L
24	OUTPUTS										
25	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
26	-554.67	1109.3318	9	1127.33	1128.08	1102.4519	6.8799	7	0.98285	1.3442	0.9872

The Log_eL is given in cell B26. Cell C26 is -2 times cell B26, and is the -2Log_eL. K is the number of parameters in any given model, and the underlying equation is =SUM(E5:E9,E12:E15). AIC is computed as the -2Log$_e$L + 2*K. AICc is the second order correction of AIC, and uses the number of study sites in the calculation. Deviance is computed as the difference between the saturated model's -2Log$_e$L and the current model's -2Log$_e$L; the lower the number the better. Remember that by definition the saturated model is a model in which the data "fit" the model perfectly. The saturated model's -2Log$_e$L is computed in the usual way (as in previous exercises) in cells N4:O21. The model we just ran had a deviance of 6.8799; it's hard to tell if this is good or not without a goodness of fit test. The Model Degrees of Freedom is the number of unique histories minus K. In a model without covariates, as long as the Model Degrees of Freedom is positive, you haven't overparameterized your model. C-hat is computed in cells J26 as Deviance divided by DF. The C-hat in this case is close to 1. C-hats larger than 1 might indicate some kind of lack of fit. The Chi-Square statistic and associated p-value are given in cells K26:L26. The Chi-square computations are provided in the orange cells L4:M19.

Click on the button labeled Model 1 to add your results to the Results Table.



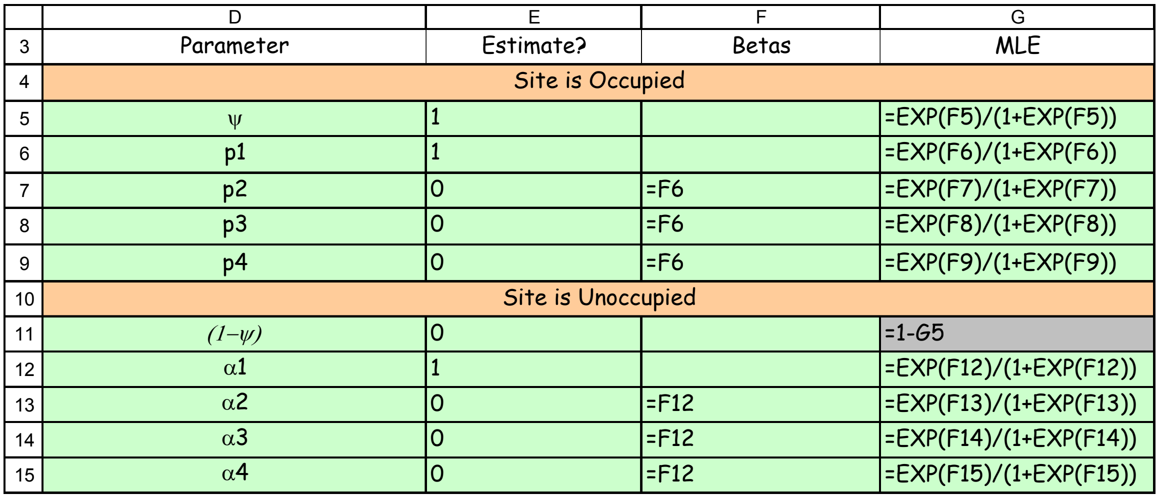
	C	D	E	F	G	H	I	J	K
29		Model	Log$_e$L	-2Log$_e$L	K	AICc	AICc	Rank	ψ hat
30	1	ψp(t)α(t)	-554.6659019	1109.331804	9	1127.331804	1128.081804	1	0.356587
31	2	ψp(.)α(.)						#N/A	
32	3	ψp(t); α = 0						#N/A	
33	4	ψp(.); a = 0						#N/A	

OK, now that you've run one model, we'll run three more: a second error model where p and alpha are constant across surveys, a model in which p varies with survey but in which there are no false positive errors, and a model in which p is constant and in which there are no false positive errors.

We will evaluate the results with model selection procedures, and will then examine the potential bias in ψ when false positives are not included in the model. The error model is not yet programmed in PRESENCE but will be in the near future.

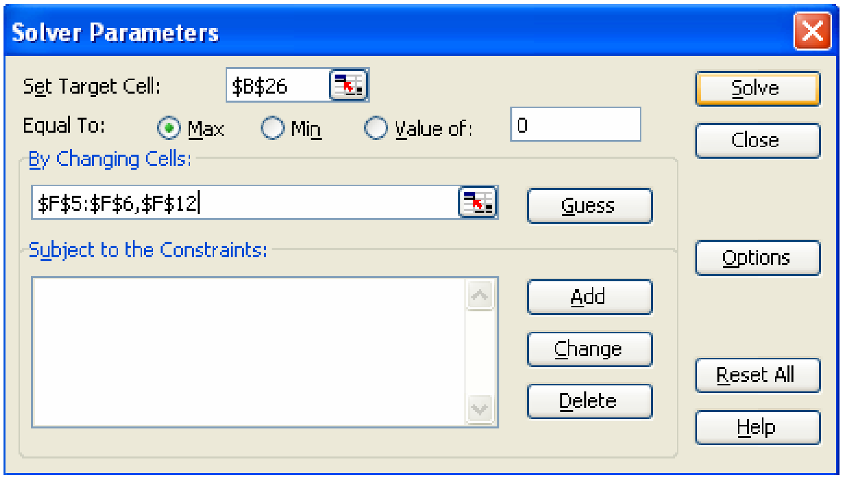
MODEL PSI, P(.), ALPHA(.)

In this model, we will estimate a single p for group 1 (sites that are occupied), will estimate a single α for group 2 (sites that are unoccupied). Think about how you would set this up in the spreadsheet.



	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1		=EXP(F5)/(1+EXP(F5))
6	p1	1		=EXP(F6)/(1+EXP(F6))
7	p2	0	=F6	=EXP(F7)/(1+EXP(F7))
8	p3	0	=F6	=EXP(F8)/(1+EXP(F8))
9	p4	0	=F6	=EXP(F9)/(1+EXP(F9))
10	Site is Unoccupied			
11	$(1-\psi)$	0		=1-G5
12	$\alpha1$	1		=EXP(F12)/(1+EXP(F12))
13	$\alpha2$	0	=F12	=EXP(F13)/(1+EXP(F13))
14	$\alpha3$	0	=F12	=EXP(F14)/(1+EXP(F14))
15	$\alpha4$	0	=F12	=EXP(F15)/(1+EXP(F15))

For group 1, we estimate ψ, and enter a 1 in cell E5. Then we estimate p_1, and enter a 1 in cell E6. Then, we force p_2, p_3, and p_4 in group 1 to be equal to p_1 and enter 0's in cells E7:E9. For group 2, we estimate α_1, so we enter a 1 in cell E12. Then we force α_2, α_3, and α_4 to be equal to α_1 for group 2. So, the total number of parameters to be estimated for this model is 3. Let's run it and see if it is more parsimonious than the previous model. Open Solver, and set cell B26 to a maximum by changing cells F5:F6,F12. Then Solve.



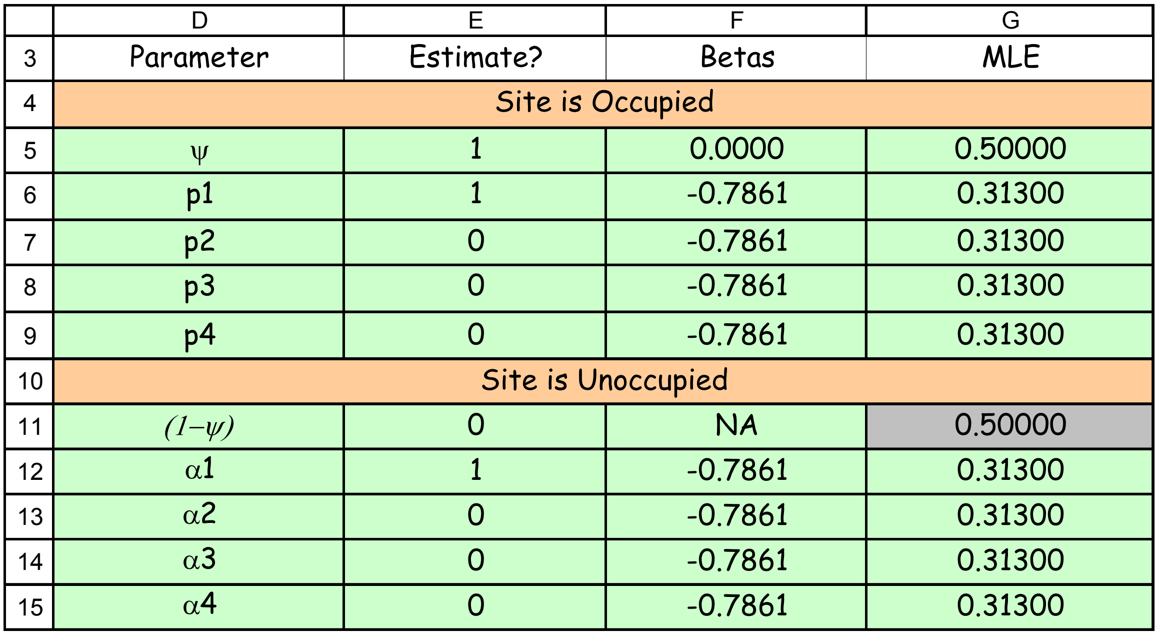
Set Target Cell: $B$26

Equal To: ● Max ○ Min ○ Value of: 0

By Changing Cells: $F$5:$F$6,$F$12

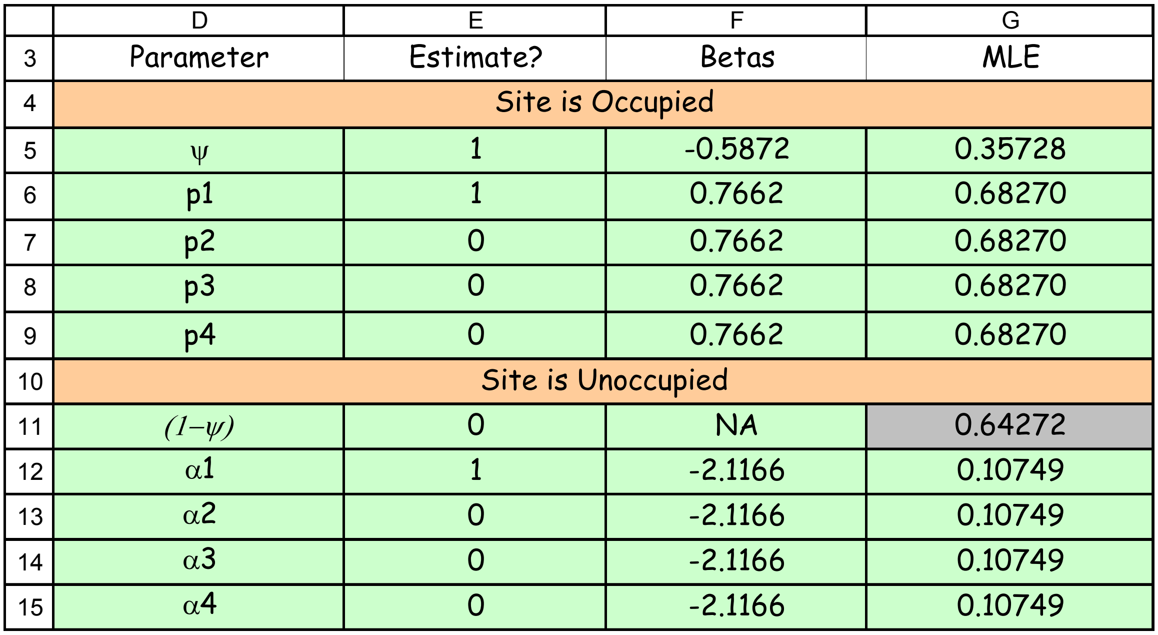
Subject to the Constraints:

Now let’s look at the output:



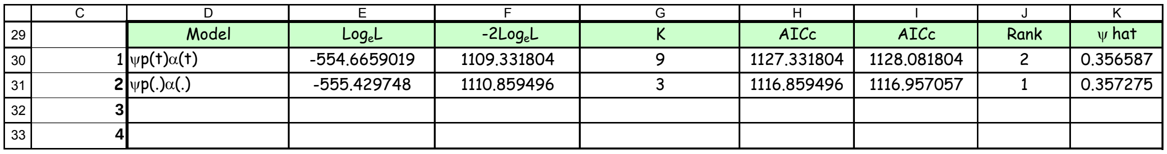
	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1	0.0000	0.50000
6	p1	1	-0.7861	0.31300
7	p2	0	-0.7861	0.31300
8	p3	0	-0.7861	0.31300
9	p4	0	-0.7861	0.31300
10	Site is Unoccupied			
11	*(1−ψ)*	0	NA	0.50000
12	α1	1	-0.7861	0.31300
13	α2	0	-0.7861	0.31300
14	α3	0	-0.7861	0.31300
15	α4	0	-0.7861	0.31300

You should notice the same problem we saw appear in the mixture model exercise, where the model generates invalid results. To overcome this, “seed” the betas with random numbers. That is, before you run Solver, enter the equation =rand() in cells F5:F6 and cell F12. Then run Solver again. Here are our new results:



	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1	-0.5872	0.35728
6	p1	1	0.7662	0.68270
7	p2	0	0.7662	0.68270
8	p3	0	0.7662	0.68270
9	p4	0	0.7662	0.68270
10	Site is Unoccupied			
11	$(1-\psi)$	0	NA	0.64272
12	α1	1	-2.1166	0.10749
13	α2	0	-2.1166	0.10749
14	α3	0	-2.1166	0.10749
15	α4	0	-2.1166	0.10749

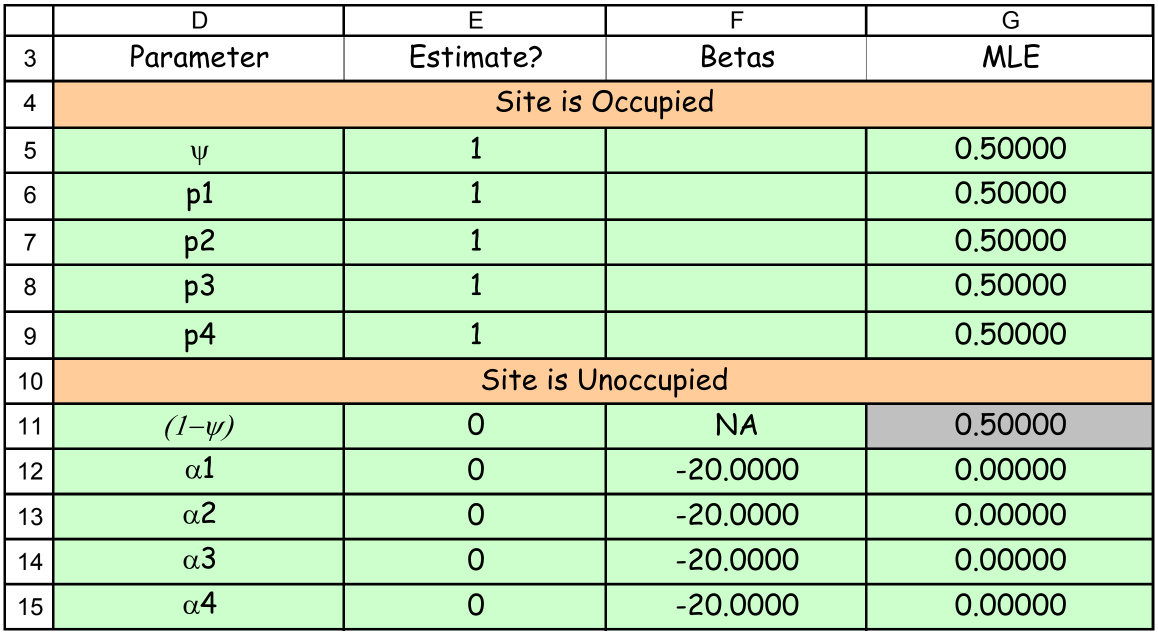
Note: your results may be "flipped." That is, you may see that the estimates for group 1 are shown in the group 2 cells and vica versa. That's OK, as long as you understand that a flip has occurred. If you run it again with new seeds, you might get the results we did….or you might get flipped results again. Just keep running the model with new random number seeds until you get the results previously shown. In this model, ψ is estimated as 0.35728, p = 0.68270, and α = 0.10749. That's a false positive error rate of around 11%, and a false-negative error rate of 1-p or 32%. Press the Model 2 button (around cell E20) to add your results to the Results Table.



	C	D	E	F	G	H	I	J	K
29		Model	Log_eL	$-2Log_eL$	K	AICc	AICc	Rank	ψ hat
30	1	$\psi p(t)\alpha(t)$	-554.6659019	1109.331804	9	1127.331804	1128.081804	2	0.356587
31	2	$\psi p(.)\alpha(.)$	-555.429748	1110.859496	3	1116.859496	1116.957057	1	0.357275
32	3								
33	4								

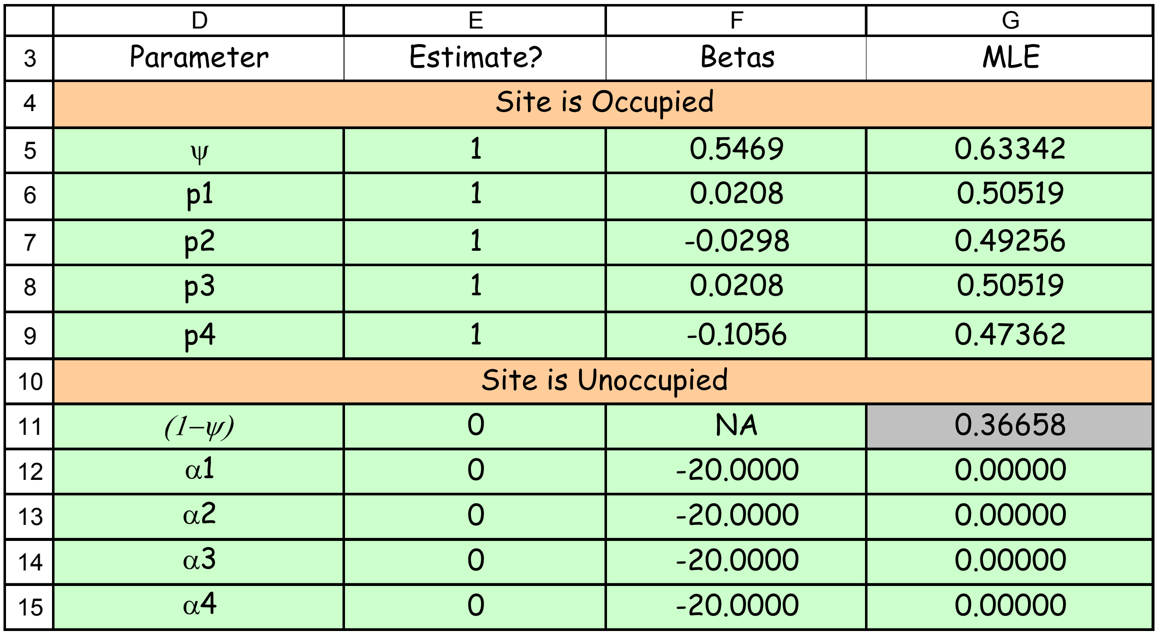
MODEL PSI,P(t); ALPHA=0

OK, two more models to go, the first of which is the standard model ψ,p(t), just like you ran in Exercise 3 and where the false positive error rate is set to 0. So, we will not be estimating any of the α parameters associated with group 2, and will force all α's to equal 0. So this model will estimate 5 total parameters. The simplest way to force the alpha's to equal 0 is to enter the number -20 in cells F12:F15, which forces the MLE's to be zero.



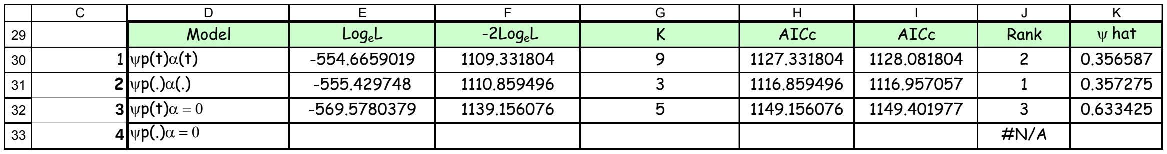
	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1		0.50000
6	p1	1		0.50000
7	p2	1		0.50000
8	p3	1		0.50000
9	p4	1		0.50000
10	Site is Unoccupied			
11	$(1-\psi)$	0	NA	0.50000
12	α1	0	-20.0000	0.00000
13	α2	0	-20.0000	0.00000
14	α3	0	-20.0000	0.00000
15	α4	0	-20.0000	0.00000

Then, just run Solver by setting cell B26 to a maximum by changing cells F5:F9. Here are our results:



	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1	0.5469	0.63342
6	p1	1	0.0208	0.50519
7	p2	1	-0.0298	0.49256
8	p3	1	0.0208	0.50519
9	p4	1	-0.1056	0.47362
10	Site is Unoccupied			
11	(1−ψ)	0	NA	0.36658
12	α1	0	-20.0000	0.00000
13	α2	0	-20.0000	0.00000
14	α3	0	-20.0000	0.00000
15	α4	0	-20.0000	0.00000

Go ahead and add the results of this model to the Results Table:

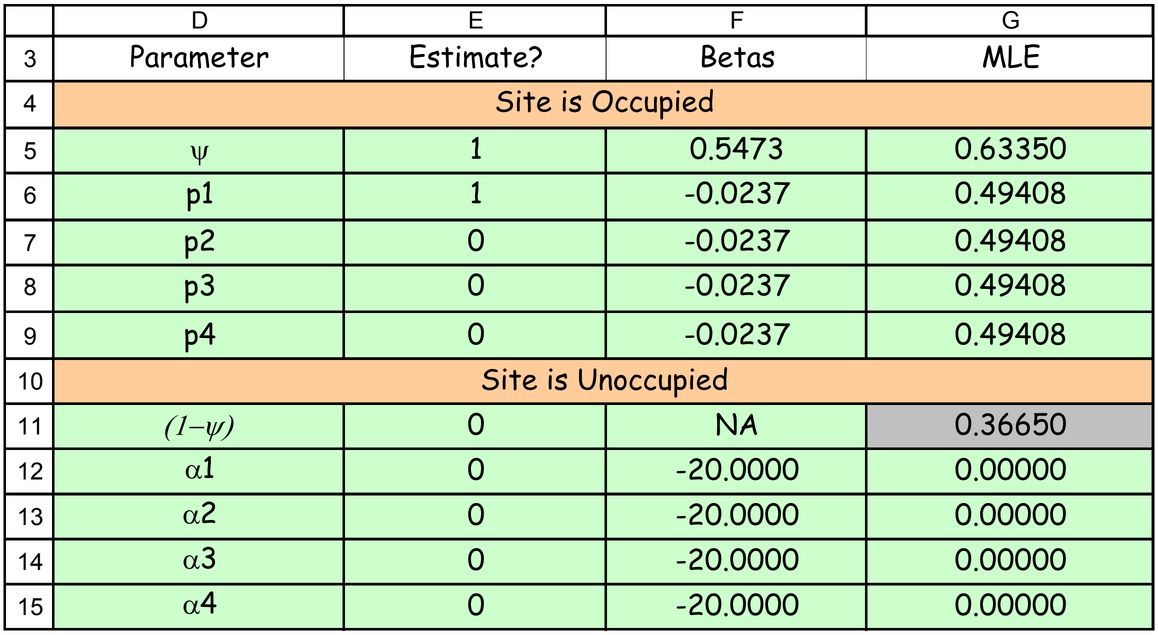


	C	D	E	F	G	H	I	J	K
29		Model	Log_eL	$-2\text{Log}_e\text{L}$	K	AICc	AICc	Rank	ψ hat
30	1	ψp(t)α(t)	-554.6659019	1109.331804	9	1127.331804	1128.081804	2	0.356587
31	2	ψp(.)α(.)	-555.429748	1110.859496	3	1116.859496	1116.957057	1	0.357275
32	3	ψp(t)α = 0	-569.5780379	1139.156076	5	1149.156076	1149.401977	3	0.633425
33	4	ψp(.)α = 0						#N/A	

You can see that the two models that account for false-positives (model 1 and model 2) are much better than model 3 (constant p and alpha). It is very enlightening to compare these models. In model 1, we estimated alpha for each survey, and estimated ψ as 0.356587. In model 2, alpha was constant and ψ was estimated as 0.357. In model 3, we ran the basic, single-season occupancy model with no false-positives, and estimated ψ as 0.633425. That's a HUGE difference in ψ estimates between the non-error model and the two error models! The data, in fact, were simulated with about a 10% false-positive error rate. If you had analyzed your data and assumed no false-positive errors, you would way over-estimate ψ (as expected – you recorded false data that indicates sites are occupied when in fact they are not).

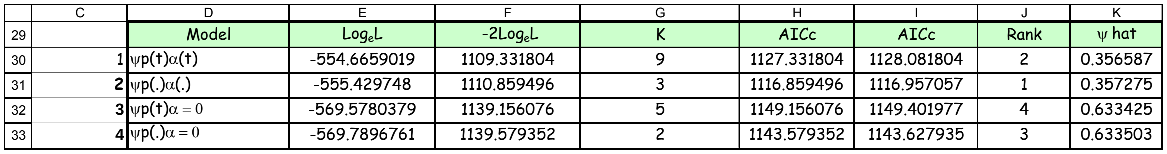
MODEL PSI,P(.); ALPHA=0

OK, the last model is the constant p model with no error rates, so you will be estimating just two parameters. Go ahead and set this model up and run it. Don't forget to force all α_i = 0 by entering -20 for their betas. Here are our results:



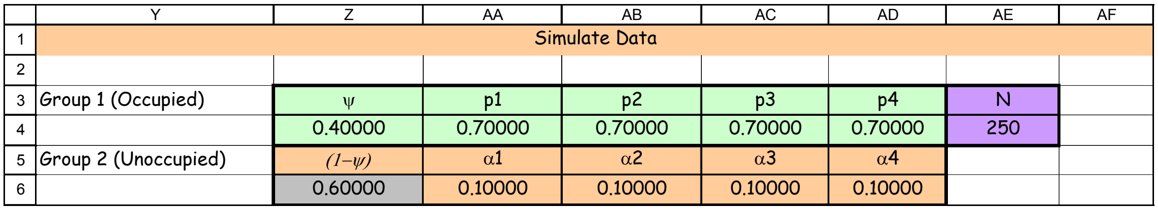
	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	Site is Occupied			
5	ψ	1	0.5473	0.63350
6	p1	1	-0.0237	0.49408
7	p2	0	-0.0237	0.49408
8	p3	0	-0.0237	0.49408
9	p4	0	-0.0237	0.49408
10	Site is Unoccupied			
11	$(1-\psi)$	0	NA	0.36650
12	α1	0	-20.0000	0.00000
13	α2	0	-20.0000	0.00000
14	α3	0	-20.0000	0.00000
15	α4	0	-20.0000	0.00000

Click on the Model 4 button to add the results to the Results Table:



	C	D	E	F	G	H	I	J	K
29		Model	Log_eL	$-2Log_eL$	K	AICc	AICc	Rank	ψ hat
30	1	ψp(t)α(t)	-554.6659019	1109.331804	9	1127.331804	1128.081804	2	0.356587
31	2	ψp(.)α(.)	-555.429748	1110.859496	3	1116.859496	1116.957057	1	0.357275
32	3	ψp(t)$\alpha = 0$	-569.5780379	1139.156076	5	1149.156076	1149.401977	4	0.633425
33	4	ψp(.)$\alpha = 0$	-569.7896761	1139.579352	2	1143.579352	1143.627935	3	0.633503

Our top-ranked model is model ψ p(.)α(.), with an AICc score of 1116.86. If you calculated the ΔAICc scores, you'd see that none of the other models support the data (ΔAICc >12). The data were in fact simulated with constant p's and constant α's, and that the occupancy rate is around 0.40.

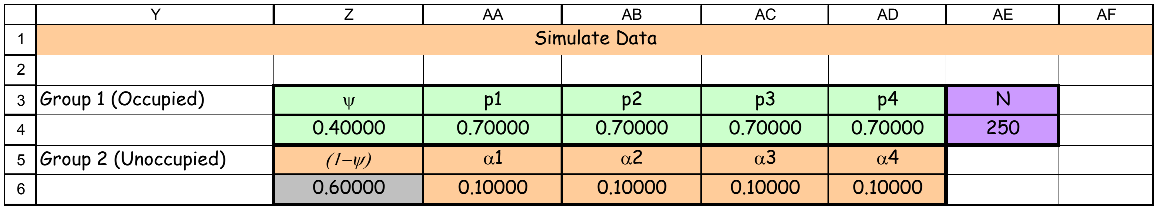


	Y	Z	AA	AB	AC	AD	AE	AF
1	Simulate Data							
2								
3	Group 1 (Occupied)	ψ	p1	p2	p3	p4	N	
4		0.40000	0.70000	0.70000	0.70000	0.70000	250	
5	Group 2 (Unoccupied)	$(1-\psi)$	α1	α2	α3	α4		
6		0.60000	0.10000	0.10000	0.10000	0.10000		

These results should really hammer home the concept that false-positives can lead to seriously biased conclusions about occupancy.

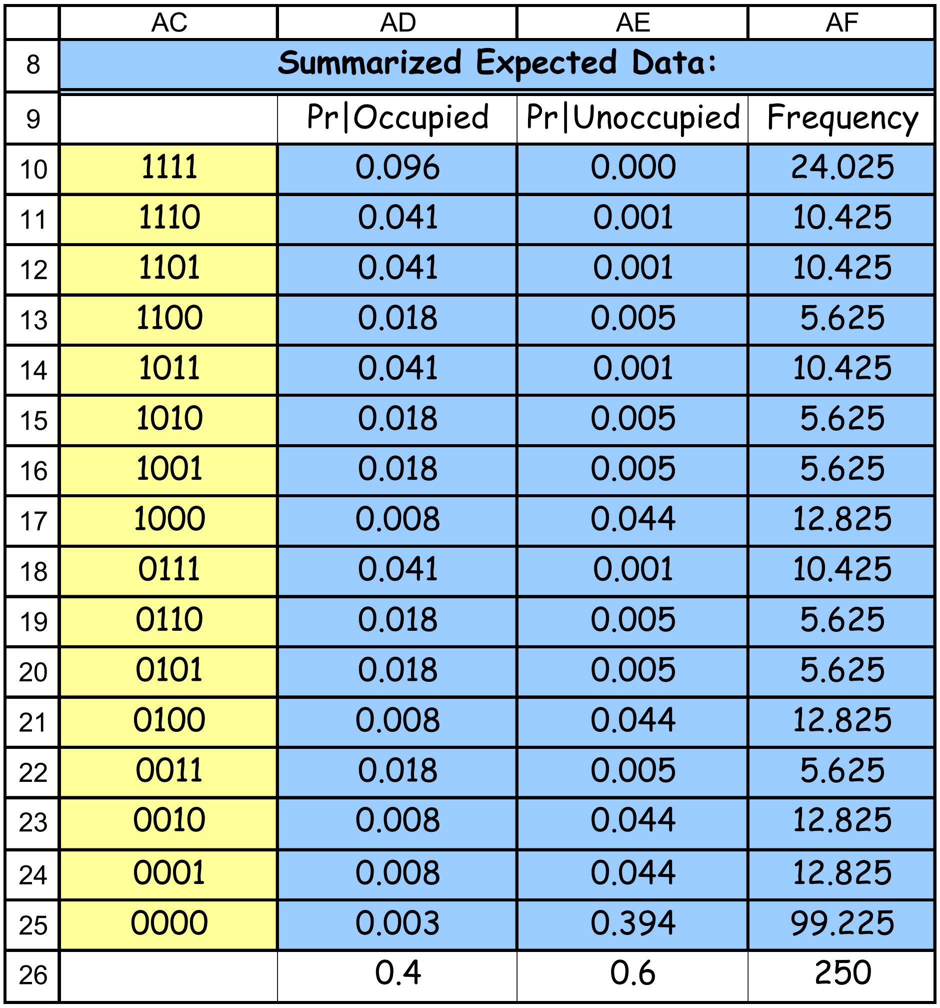
SIMULATING ERROR DATA

Before we finish, we want to demonstrate how the data were simulated for this exercise. We already mentioned that the data were simulated where ψ = 0.4, p = 0.7, and α = 0.10. You can simulate any estimates you'd like. Start by entering the total number of sites in cell AE4, and then enter the occupancy rate in cell Z4. Note that you don't have to enter the vacancy rate (cell Z6); it is grayed out because it is computed. Then, for sites that are occupied, enter values for p_1, p_2, p_3, p_4 in cells AA4:AD4. For sites that are empty, enter α_1, α_2, α_3, and α_4 in cells AA6:AD6.



	Y	Z	AA	AB	AC	AD	AE	AF
1	Simulate Data							
2								
3	Group 1 (Occupied)	ψ	p1	p2	p3	p4	N	
4		0.40000	0.70000	0.70000	0.70000	0.70000	250	
5	Group 2 (Unoccupied)	$(1-\psi)$	α1	α2	α3	α4		
6		0.60000	0.10000	0.10000	0.10000	0.10000		

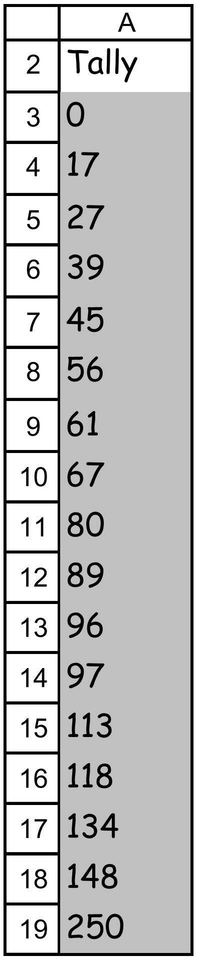
As with the other spreadsheet exercises, we will simulate data in two ways: by expectation and with stochasticity. The expected data are simulated in cells AC8:AF26. As we did in the spreadsheet exercise, we enter encounter history probabilities for each group separately, and then add the two probabilities together. Cells AD10:AD25 give the encounter history probabilities for sites that are truly occupied (group 1), while cells AE10:AE25 give the encounter history probabilities for sites that are truly empty (group 2). The expected frequency of each history is computed in cells AF10:AF25 as N (cell AE4) times the sum of group 1 + group 2's encounter history probability.



	AC	AD	AE	AF
8	**Summarized Expected Data:**			
9		Pr|Occupied	Pr|Unoccupied	Frequency
10	1111	0.096	0.000	24.025
11	1110	0.041	0.001	10.425
12	1101	0.041	0.001	10.425
13	1100	0.018	0.005	5.625
14	1011	0.041	0.001	10.425
15	1010	0.018	0.005	5.625
16	1001	0.018	0.005	5.625
17	1000	0.008	0.044	12.825
18	0111	0.041	0.001	10.425
19	0110	0.018	0.005	5.625
20	0101	0.018	0.005	5.625
21	0100	0.008	0.044	12.825
22	0011	0.018	0.005	5.625
23	0010	0.008	0.044	12.825
24	0001	0.008	0.044	12.825
25	0000	0.003	0.394	99.225
26		0.4	0.6	250

The stochastic data are created in a similar way, except that we use random numbers to assign each site to either group 1 or group 2. For sites in group 1, we also use random numbers associated with p1 – p4 that determine whether the species was detected or not on a given survey. For sites in group 2, we use random numbers associated with α1 – α4 that determine whether the species was falsely recorded as being present on a survey. This should be fairly straight-forward to you by now, so take some time clicking on the formulae in cells AJ29:AL29 to see if you can understand how the encounter history was generated for site 1.

PRESENCE INPUT FILES



	A
2	Tally
3	0
4	17
5	27
6	39
7	45
8	56
9	61
10	67
11	80
12	89
13	96
14	97
15	113
16	118
17	134
18	148
19	250

The histories and corresponding frequencies given in cells B4:B11 cannot be input directly into PRESENCE (most users of PRESENCE include covariates in the analysis, so the input files are set up on a site-by-site basis). So, we've entered some formulae in columns Q:V to convert the summarized data to site-specific data. But before we cover the equations, first look at cells A2:A19, which are shaded grey on the spreadsheet. These cells are a running tally of the total number of sites in the study. Beginning with the first history (1111), the cell A4's formula counts the number of sites that are 1111. The next cell (cell A5) counts the number of 1110 sites + the 1111 sites. The next cell (cell A6) counts the number of 1101, 1110, and 1111 sites, and so on. We will use this running tally to create PRESENCE input files.

Now let's turn our attention to columns Q:V. In column Q, the sites are listed from 1 to 250 down the column. In column R, we assign a history to each site, using the tally in cells A3:A19. Click on cell R4. The equation there is =LOOKUP(Q4-1,$A$3:$A$19,$B$4:$B$19). The function looks up the value in Q4 (the site number) minus 1 in the tally column (A3:A19), and then returns the corresponding history listed in cells B4:B19. Because the lookup vector (the tally) is sorted in ascending order, this equation "works" for our purposes because the LOOKUP function doesn't need to find an exact match. Take a look again at cells B4:C19. Notice that there are 17 sites with a 1111 history, 10 sites with a 1110 history, 12 sites with a 1101 history, and so on. When the lookup function in column R is copied down the column, the result is that the first 17 sites are given a 1111 history, the next 10 sites are given a 1110 history, the next 12 sites are given a 1101 history, and so on. Given the histories in column R, columns S:V simply split each history into survey-specific results (using LEFT, RIGHT, and MID functions). When it's time to create a PRESENCE input file, simply copy cells S4:V253 and paste them into the PRESENCE datasheet.