Logistic Regression
For Matched Pairs Data

Eric Meredith
Department of Mathematical Sciences
Montana State University

May 11, 2007

A writing project submitted in partial fulfillment
Of the requirements for the degree

Masters of Sciences in Environmental and Ecological Statistics

Introduction

Variations in climate affect the behaviors of species in their environments. The Upper Madison River elk herd in Yellowstone National Park undergoes many responses of behavior as environmental conditions change. This non-migratory herd of elk stays in the same general area year round, including the often harsh winter season.

This study is a small part of a bigger project that is aimed at investigating the impacts of changing environmental impacts on elk behavior. The Upper Madison River area receives a substantial amount of snowfall each year that generally begins during October. The arrival of snow each year influences the behaviors of elk in the valley. The objective of this study is to examine the relationships between snow dynamics and elk behavior. The impact on elk behavior due to the reintroduction of wolves, which occurred during the time period of this study, was also of interest.

The method used to analyze this dataset will be matched case-control logistic regression. This method will allow the comparison of observed use of the area to what would have been used under random selection.

First the specifics of logistic regression will be discussed which will lead to a discussion of matched case-control logistic regression. Following this, a one to one matched case example will be presented to show matched case-control logistic regression in use. Then the analysis of the Upper Madison River elk herd will be used to show an example of 1-M matched case-control with logistic regression.

Background

A binary response variable, Y, takes on a value of 1 for a "success" and a value of 0 for a "failure". An explanatory variable, X, leads to π(x) defined as the probability of success given x.

Thus the probability of success, $\pi(x) = P(Y=1|x)$, is a function of one or more covariates that may be a mixture of categorical and numerical variables. For notational convenience throughout the rest of the paper all models will assume a single predictor, however all these models also hold for multiple predictors.

A linear probability model for a binary response variable Y assumes

$$\pi(x) = \alpha + \beta x \qquad (1.1)$$

The linear model is not ideal. The assumption of a linear relationship between a covariate x and probability is a strong one and unlikely to be true in general. For example, the linear model may yield estimated probabilities outside the range of zero to one, especially for extreme values of x. The linear model is rarely used in practice because of these and other problems.

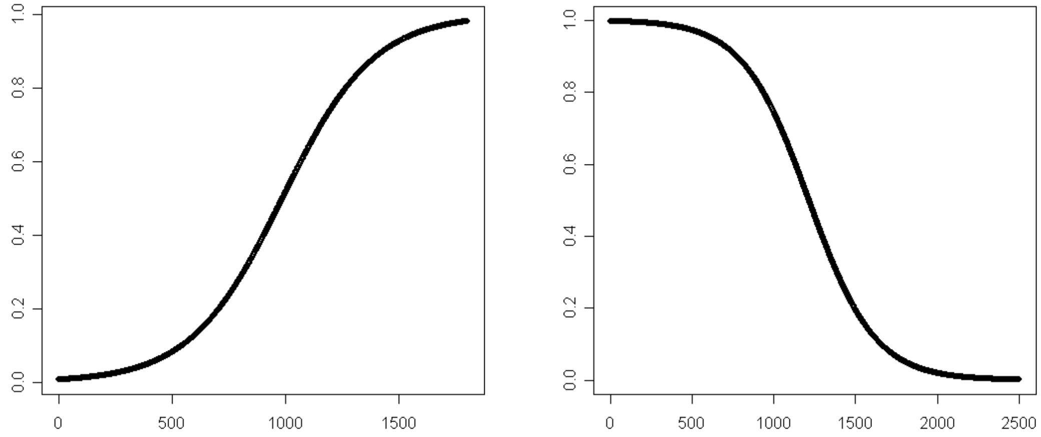
A number of alternatives to the linear model (1.1) exist. A more appropriate model is the logistic regression model.

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad (1.2)$$

The logistic regression model has a general S shape pattern (Figure 1) that is seen more commonly with binary responses.

As β moves further from zero, the rate of change increases. Logistic regression models also follow the rules of probability and may never be outside of the range from zero to one.

<u>Figure 1 – Shapes of logistic regression line dependent on β</u>
(y-axis is the probability of a success)



β > 0 β < 0

Understanding the interpretation of the parameters of the logistic regression model requires the odds. The odds of an event occurring are defined as

$$\text{odds} = \frac{\pi(x)}{1-\pi(x)} \qquad (1.3)$$

If π(x) = .8 then the odds of success equal .8/.2 = 4. This is interpreted to mean a success is four times as likely to occur as compared to a failure.

An odds ratio is defined as

$$\theta = \frac{\text{odds1}}{\text{odds2}} = \frac{\pi(x=x_1)}{1-\pi(x=x_1)}\bigg/\frac{\pi(x=x_2)}{1-\pi(x=x_2)} \qquad (1.4)$$

The odds ratio compares the odds of success when $x = x_1$ to the odds when $x = x_2$. The odds ratios are generally interpreted with *x_2=x_1+1* but can be found for other values. When the odds ratio is one then the odds are equal. When it is greater than one, then the odds of success when $x = x_1$ is more likely than the odds when $x = x_2$. When the odds ratio is less than one, the odds of success when $x = x_1$ is less likely than the odds when $x = x_2$.

The *logit* function can be used to arrange the logistic regression model into a form that is more easily interpretable. The *logit* is a log of the odds and yields a linear function of the explanatory variables similar to equation (1.1).

$$\operatorname{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x \qquad (1.5)$$

From (1.5) it can be seen that

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x} \qquad (1.6)$$

Note that

$$\frac{\pi(x+1)}{1-\pi(x+1)} \Bigg/ \frac{\pi(x)}{1-\pi(x)} = e^{\beta} \qquad (1.7)$$

Thus the odds of success at x + 1 are e^{β} times the odds of success at x. The intercept term, α, is the odds of success when x = 0.

For k predictor variables the logistic regression model in linear form (1.5) is

$$\operatorname{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad (1.8)$$

Each x_i is a different explanatory variable of interest. This does not change the interpretation of the parameters, except there are now several odds ratios to estimate at once.

One advantage of logistic regression is it may be used in case-control studies. Case-control studies are examples of retrospective studies in that samples are taken after the events of interest (success or failure) have occurred. In a case-control study, separate samples of cases (Y=1) and controls (Y=0) are taken and the values of the predictor variables ($x_i, i = 1, \cdots, p$) are observed on the selected units. There is evidence of an association if the distribution of predictors is different in cases and controls.

The proportion of cases in a case control sample will generally not be equal to the proportion in the population and it is not possible to estimate the probability of a case in the population. However it is possible to use logistic regression to draw inferences about odds ratios with case-control data.

In case-control studies it is sometimes desirable to match each case with one or more controls. The typical reason for matching in a case control study is to control for variability due to predictors that are not of much interest. The matching variable is typically one whose effects are likely to be confounded with other predictors. Care must be taken in the selection of matching variables. If matching is done on a variable of interest then all of the cases are matched with controls on that variable and in general it will not be possible to come to any conclusions regarding that variable.

It is possible to add interaction terms involving the matching variable and other predictors to investigate the effects of predictors across the different values of the matching variable. This is one of the few times in statistical models it is appropriate to add an interaction term when both variables are not already included in the model as a main effect.

Matched pairs in a case control study is most commonly done as a 1 – 1 design where one case is matched to one control. This method may also be used in a 1 – M design, in which one case is matched to M controls.

In the case of matching on a variable, the earlier mentioned models are going to change slightly for logistic regression. The new model in linear form is

$$\log\text{it}[\pi_j(x_{ij})] = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_k x_{kij} \qquad (1.9)$$

where x_{hij} is the value of the *h*th explanatory variable for the *i*th individual in the *j*th matched set. The α_j will vary for each of the j matched sets. However, even with a different α_j for each matched set this does not have an effect on the odds ratios because the α terms cancel out of the odds ratio calculations(see equation 1.11).

The parameters of the logistic regression models (1.8) are estimated using the method of maximum likelihood. For the matched case-control scenario, the method of maximum likelihood must be modified (see Hosmer and Lemeshow, 2000, pages 225-226). For the 1-1 matched case-control scenario, logistic regression software can be used with the appropriate modifications. For 1-M matched case-control, special software is needed. In SAS 1-M matched data can be fit using PROC PHREG.

Additional details on logistic regression in general, and matched case-control logistic regression in particular can be found in other literature (Agresti, 2002) (Collett, 2003) (Hosmer and Lemeshow, 2000).

1 – 1 Matched Pairs Logistic Regression Example

As an example, a dataset consisting of human mothers that had low birth weight children and matching them with mothers who did not have a low birth weight baby will be analyzed. This dataset consisted of 56 case-control pairs matched so that each case-control pair has the same age. The variables available are race (RACE), smoking status (SMOKE), presence of hypertension (HT), presence of uterine irritability(UI), presence of previous pre-term delivery (PTD) and the weight of the mother at the last menstrual period (LWT). A binary variable (1 or 0) is assigned for smoking status (yes/no), presence/absence of hypertension, presence/absence of uterine irritability, and presence/absence of pre-term delivery with a one indicating the presence of the condition. The data set was obtained from Hosmer and Lemeshow(2000).

As an example, the results of fitting a model with only one numerical predictor variable, the weight of the mother at the last menstrual period, are presented in Table 1.

Table 1 - Parameter Estimate for Low weight 1-1 Matched Pairs Model

Variable	Parameter Estimate	Standard Error	Pr > ChiSq
LWT	-0.00937	0.00617	0.1284

The estimate for β is $\hat{\beta} = -0.0094$ (SAS computer code in appendix A, part 1) with $e^{-.0094}$ = 0.991. The results can be interpreted any number of different ways. One interpretation for this parameter is for every one pound increase in the weight, at the last menstrual period, mothers are .991 times more likely to have a low weight baby. A second interpretation is the ratio of the odds of a low weight baby at a weight of $x+1$ pounds to the odds at a weight of x pounds is 0.991 for any value of x. It can also be stated that we estimate each one pound increase in weight leads to an approximate 1%

decrease in the odds of a low weight baby. All these interpretations suggest the chances of having a low weight baby decrease as the mother's weight increases.

A value of e^{β} near one means there is no relationship between the odds are the predictor variable. The estimate of 0.991 in this example is very close to one for a couple of reasons. First of all it indicates the change in the odds based on a one pound increase in weight.

However trying to interpret the odds based on a one pound increase in a mother's weight may not make sense. For example, it may make more sense to interpret the results in terms of ten pound changes in body weight. It is estimated the odds of a low weight baby at a weight of $x+10$ pounds is $e^{10(-0.0094)}$ = 0.091 times the odds of a low weight baby at a weight of x pounds. To find the new standard error for a ten pound increase, the standard error for a one pound increase is multiplied by ten also.

It is also appropriate to find large sample confidence intervals for β and e^{β}. To find an approximate (large sample) confidence interval for β_i the following equation is used:

$$\beta_i \pm z_{1-\alpha/2} * SE[\beta_i] \qquad (1.10)$$

The 95% confidence interval for the estimate of β in this data is -0.00937 ± 1.96 * 0.00617 = (-0.0215, 0.00272). To find the 95% confidence interval for e^{β}, exponentiate the endpoints, which yields $(e^{-0.0215}, e^{0.00272})$ = (.978, 1.003). As mentioned above it may not be relevant to interpret results in terms of a one unit increase in the value of a predictor variable. It is easy to modify the confidence interval formulas to account for other changes. For example, the confidence interval for impact for a ten pound increase in weight can be found by (-0.0215*10 , 0.00272*10) = (-0.215, .0272).

Finding the 95% confidence interval for $e^{10\beta}$ yields ($e^{-0.215}$, $e^{0.0272}$) = (.807, 1.03). This confidence interval suggests there could be as much as a 19 percent decrease in the risk of a low weight baby to a three percent increase in a low weight baby with a ten pound increase in the mother's weight.

The interpretation for a categorical predictor variable is somewhat different than that of a numerical predictor variable. The results of a logistic regression model with smoking status (SMOKE) as the predictor are shown in Table 2.

Table 2 - Parameter Estimate for Smoke 1-1 Matched Pairs Model

Variable	Parameter Estimate	Standard Error	Pr > ChiSq
SMOKE	1.0116	0.41286	0.0143

This binary predictor variable is coded as a 1 if the mother is a smoker and 0 if she is not. The estimate for β is 1.0116 (SAS computer code in appendix A, part 1). The estimate for e^{β} is $e^{1.011}$ = 2.75. The interpretation for this estimate is a mother that has a history of smoking during the pregnancy is estimated to be 2.75 times more likely to have a low weight baby than a mother who does not have a history of smoking during the pregnancy.

The data set for low weight babies has multiple predictors in it. Before the multiple logistic regression model is run the data set must be prepared. The RACE variable is a nominal categorical variable with three levels of 1(White), 2(Black), and 3(Other).

Due to the nominal categorical variable, separate binary indicator variables may be needed to be made for the levels of black and other, RACE2 and RACE3 respectively, depending on the software being used. A separate variable is not needed for the level of white, since the person will be assumed to be white unless otherwise noted by the new indicator variables. The logistic regression model for multiple predictors is run with LWT, SMOKE, RACE2, RACE3, PTD, HT, and UI which gives the results for the estimates of β shown in Table 3.

Table 3 - Parameter Estimates for 1-1 Matched Pairs Model

Variable	Parameter Estimate	Standard Error	Pr > ChiSq
LWT	-0.01838	0.01008	0.0683
SMOKE	1.40066	0.62784	0.0257
RACE2	0.57136	0.68964	0.4074
RACE3	-0.02531	0.69920	0.9711
PTD	1.80801	0.78865	0.0219
HT	2.36115	1.08613	0.0297
UI	0.40193	0.69616	0.0440

A negative parameter estimate means the odds will decrease as that variable increases and a positive estimate means the odds will increase as that variable increases. Notice the parameter estimates for LWT and SMOKE have changed slightly with the addition of the other variables. The interpretations follow the same format as before. For example, for PTD, the estimate for e^{β_5} is $e^{1.808}$ = 6.10. This can be interpreted as a mother having a pre-term delivery is estimated to be 6.10 times more likely to have a low weight baby as a mother that does not have a pre-term delivery. This interpretation assumes all other variables in the model are the same for these two mothers.

Estimates for combinations of predictor variables can be found. For example, using the estimates, the odds ratio of a low weight baby from a black woman with the presence of hypertension as compared to that of a white woman without the presence of hypertension is estimated by

$$\theta = \frac{\text{odds1}}{\text{odds2}} = \frac{\exp(\alpha + \beta_3 + \beta_6)}{\exp(\alpha)} = \exp(\beta_3 + \beta_6) \qquad (1.11)$$

Using the estimates found earlier then the estimated odds ratio for $e^{\beta_3+\beta_6}$ is $e^{0.57136+2.36115}$=18.77. The interpretation is a black mother with hypertension is estimated to be 18.77 times more likely to have a low weight baby than a white mother without a presence of hypertension. Whichever odds ratios are of interest may be found using different combinations of the parameter estimates.

1 – M Matched Pairs Logistic Regression Example – Yellowstone Elk Herd

Matched pairs logistic regression may also be done by matching M controls to one case. An example of this is the Upper Madison River elk herd data mentioned before.

Each winter the valley receives different amounts of snowfall. The amount of snow is recorded using snow water equivalent (SWE). This is a better measure than snow depth alone because SWE also takes into consideration the density of the snow on the ground.

Elk location was determined by randomly selecting female elk (25-40 individuals) to be fitted with a radio collar. The location of these elk was randomly checked during the winter. Locations of elk use were recorded and habitat variables of interest were measured at each location. Each location was matched with twenty randomly selected locations and the same set of habitat variables were measured at each random location. The data were collected over a period of time prior to and after wolf reintroduction (Messer, 2003). It is then natural to compare the impact of wolves on habitat selection by elk.

Logistic regression is used with the binary response variable (1 for observed elk use, 0 for a randomly chosen location) along with the explanatory variables, including SWE to explore elk habitat use and habitat predictor variables of interest.

The data set was split into three different sets of data split up by time: prewolf (1991-1998), transition period (1998-2002), and an established period (2002-2006). The transition period is needed because the elk were still adjusting to wolves being in the area and not all elk in the study area had been introduced to wolves.

With all of the variables available for this data set the first step is to determine which variables should be used in the model. Each of the variables used in the models were standardized by subtracting the difference of the highest and lowest values and dividing all of this by half the range. This scales all values to be between -1 and 1. This scaling procedure allows for the direct comparison of the coefficient values in the model.

Twenty four different models were fit with matched pairs logistic regression (SAS PROC PHREG computer code provided in Appendix A, part 3) and were compared to each other using Akaike's Information Criteria (AIC) values (see Appendix B for results). The best fit model according to AIC was the same model for the 3 different time periods. This model included the variables SWEA (local snow water equivalent), SNHA (local snow heterogeneity), HBT (habitat type), ELV (elevation), and SRI (Solar Radiation Index). All of these variables are numerical variables except habitat. The baseline level for HBT was geothermal areas with the other factors being burned forest (BF), unburned forest (UB), and meadow (M). The parameter estimates for the β's for each of the time periods is given in Table 4.

Table 4 – Parameter Estimates for each of three time periods

	Pre-Wolf		Transition Period		Established Wolf	
	Parameter Estimates	Standard Error	Parameter Estimates	Standard Error	Parameter Estimates	Standard Error
SWEA	-2.83	0.12	-3.36	0.25	-4.96	0.31
SNHA	5.80	0.24	7.16	0.51	7.74	0.64
HBT(BF)	-0.58	0.06	-0.65	0.11	-0.74	0.11
HBT(UB)	-0.38	0.05	-0.32	0.09	-0.28	0.10
HBT(M)	-0.46	0.06	-0.59	0.11	-0.77	0.12
ELV	-2.06	0.06	-2.21	0.09	-2.87	0.10
SRI	-0.28	0.10	-0.46	0.15	-1.20	0.15

Since each of the models is measuring the same thing, only during different time periods, we can directly compare the coefficients from one time period to another. This is important because we are then able to compare how the elk responded due to the re-introduction of wolves.

The interpretation of the odds ratios is a little more difficult for this example. All the numerical variables in the model were scaled from -1 to 1 and because of this the odds ratios are interpreted a little differently. For example, for the pre-wolf data, SWEA has a parameter estimate of -2.83 with $e^{-2.83} = .059$. The interpretation is for every one unit increase in SWEA elk is estimated to be only 0.059 times as likely to select that area. However a one unit increase under the scaling procedure is half the range of that variable. For this data set there is more interest in being able to compare the parameter estimates to each other. This way it is easier to see which variable has more strength in determining the location of an elk.

Looking at just the pre-wolf estimates in Table 4, it appears from the SWEA estimate the elk are more likely to be located in areas with lower amounts of snow. The SNHA estimate suggests this is true especially in areas that have snow heterogeneity.

There are four levels of the habitat variable: thermal, burned forest, unburned forest and meadow. We only have parameter estimates for three of them(BF, UN and M). This makes thermal areas the reference for parameter estimates of the other three. Since all the estimates for the habitat variables are negative, the elk seem to prefer the thermal areas over all of the other areas: burned forest, unburned forest and meadow. Because of the fact the elk seem to prefer to be in areas with lower SWEA values, then it is expected they would also want to be in lower elevation areas and this is backed up by the parameter estimate for ELV. However it must be mentioned there are could be multicollinearity problems between ELV and SWEA because as the elevation decreases the amount of snow is going to generally decrease. The correlation between ELV and SWEA is only .40 so multicollinearity is not a big concern for this model.

Approximate 95% confidence intervals for all the parameter are shown in Table 5.

Table 5 – 95% Confidence Intervals of the parameters for each of 3 time periods

	Pre-Wolf	Transition	Established
	Standard Error	Parameter Estimates	Parameter Estimates
SWEA	(-3.06, -2.59)	(-3.85, -2.87)	(-5.57, -4.35)
SNHA	(5.33, 6.27)	(6.16, 8.16)	(6.49, 8.99)
HBT(BF)	(-.70, -.46)	(-.87, -.43)	(-.96, -.52)
HBT(UB)	(-.48, -.28)	(-.50, -.14)	(-.48, -.08)
HBT(M)	(-.58, -.34)	(-.81, -.37)	(-1.01, -.53)
ELV	(-2.18, -1.94)	(-2.39, -2.03)	(-3.07, -2.67)
SRI	(-.48, -.08)	(-.75, -.17)	(-1.49, -.91)

Comparing the 95% confidence intervals of each variable it can be determined if the coefficients are statistically significant from one time period to another. A comparison of the coefficients for SWEA provides information on the possible impacts of wolf reintroduction on elk behavior. The difference from the pre-wolf to the transition time periods is not significantly different, however the coefficient from the established time period is significantly different from either of the two earlier periods. The negative coefficients tell us there is significant evidence to suggest elk are locating in areas of less snow with the introduction of wolves. The SNHA coefficient of the established time period is also significantly different from the coefficient for the pre-wolf time period. Since this coefficient is positive there is also strong evidence to suggest the elk are locating in areas that have higher snow heterogeneity. It is also noteworthy to notice the coefficients for habitat type are not statistically different from each other in any of the time periods. So the elk do not seem to have drastically changed the type of habitat they choose to inhabit because of the introduction of wolves. However it appears there may be some degree of movement away from the meadow areas that may need to be investigated further.

Conclusion

Matched pairs logistic regression can be the appropriate way to run the statistical analysis, given the appropriate dataset. When two samples are statistically dependent on each other and can naturally be matched together to form cases and controls the dataset could be used for matched pairs logistic regression. We use matching to control the variability due to predictors we are not generally interested in. For example the elk study had 20 random points that were matched by date to the one case that was the actual location of an elk.

The Madison elk herd study is an example of a dataset that can be analyzed using matched pairs logistic regression. The results of the study provide evidence the elk in general tend to use areas of less snow during the winter. This tendency was more pronounced after the reintroduction of wolves. This may be due to the ability of elk to run away from wolves easier in less snow. Elk may choose areas of high snow heterogeneity so they can expend less energy in finding food. The reason for the increase in the use of the high snow heterogeneity areas following reintroduction of wolves may be that elk can run away faster and get to areas of no or less snow.

Literature Cited

Agresti A. , 1996, An Introduction to Categorical Data Analysis, John Wiley and Sons, Inc, New York, New York, USA

Agrest A. , 2002, Categorical Data Analysis, 2nd edition, John Wiley and Sons Inc, New York, New York, USA

Collett, D. , 2003, Modeling Binary Data, 2nd edition, Chapman and Hall Press LLC, Boca Raton, Florida, USA

Hosmer D. , S Lemeshow, 2000, Applied Logistic Regression, 2nd edition, John Wiley and Sons Inc, New York, New York, USA

Messer A. , 2003, Identifying Large Herbivore Distribution Mechanisms Though Application of Fine-Scale Snow Modeling

Stokes M. E. , C. S. Davis, and G. G. Koch, 2000, Categorical Data Analysis Using the SAS system, SAS Institute Inc, Cary, North Carolina, USA

Appendix A

Part 1

SAS computer code for 1 – 1 matched case-control logistic regression for low weight birth mothers. This includes the code for all 3 models in the paper.

```
data try;
infile ‘c:\lowwtbaby.txt’;
input ID LOW AGE LWT RACE $ SMOKE PTD HT UI;
RACE2=0;
RACE3=0;
IF RACE='2' THEN RACE2=1;
IF RACE='3' Then Race3=1;
case=2-LOW;

proc phreg;
strata ID;
model case = SMOKE;


proc phreg;
strata ID;
model case = LWT;

proc phreg;
strata ID;
model case = LWT SMOKE RACE2 RACE3 PTD HT UI;
run;
```

Part 2

SAS computer code for finding the midpoints and ranges to scale the elk data from -1 to 1 for each explanatory variable

```
dm 'log;clear;out;clear;';
data pre;
infile 'PREWOLF.CSV' delimiter=',';
input east north date mmddyy8. elv elkid randid slope aspect GHF MAP sri LULC
SWE_100 Stdev_100 StudySWE StudyStdev ernd bf uf md;
Format date mmddyy8.;

data post;
infile 'PostWolfTrans.csv' delimiter=',';
input east north date mmddyy8. elv elkid randid slope aspect GHF MAP sri LULC
SWE_100 Stdev_100 StudySWE StudyStdev ernd bf uf md;
Format date mmddyy8.;
```

```
data est;
infile 'PostWolfEstab.csv' delimiter=',';
input east north date mmddyy8. elv elkid randid slope aspect GHF MAP sri LULC
SWE_100 Stdev_100 StudySWE StudyStdev ernd bf uf md;
Format date mmddyy8.;

data pooled; Set pre post est;

If SWE_100<0 Then delete;
If Stdev_100<0 Then delete;

swea=SWE_100;
snha=Stdev_100;
sweasnha=swea*snha;
swel=StudySWE;
snhl=StudyStdev;

laswea=swea*swel;
lasnha=snha*swel;
lhswea=swea*snhl;
lhsnha=snha*snhl;

lhelv=snhl*elv;
lhsri=snhl*sri;
lhbf=bf*snhl;lhuf=uf*snhl;lhmd=md*snhl;
snbf=bf*swea;snuf=uf*swea;snmd=md*swea;
sweasnha=swea*snha;
shbf=bf*snha;shuf=uf*snha;shmd=md*snha;
elvbf=bf*elv;elvuf=elv*uf;elvmd=elv*md;
sribf=bf*sri;sriuf=uf*sri;srimd=md*sri;

Proc Means Data=pooled;
Var swea snha sweasnha sri elv lhelv;
Output out=center
Min= mswea msnha msweasnha msri melv mlhelv
Max= xswea xsnha xsweasnha xsri xelv xlhelv;
Title 'Center and Scale';

Proc Print Data=center;

Data center; Set center;
midswea=(xswea+mswea)/2;
Rswea=xswea-mswea;
midsnha=(xsnha+msnha)/2;
Rsnha=xsnha-msnha;
```

```
midsweasnha=(xsweasnha+msweasnha)/2;
Rsweasnha=xsweasnha-msweasnha;
midsri=(xsri+msri)/2;
Rsri=xsri-msri;
midelv=(xelv+melv)/2;
Relv=xelv-melv;
midlhelv=(xlhelv+mlhelv)/2;
Rlhelv=xlhelv-mlhelv;

Proc Print Data=center;

run;
```

Part 3

SAS computer code for 1-M matched case control logistic regression of the elk data. This includes scaling, coding variable names for interactions, and other explanatory variables. All models are included in the code for the prewolf data. The only change for the transition and established time periods is the infile and the data specified for each model.

```
dm 'log;clear;out;clear;';
data pre;
infile 'PREWOLF.CSV' delimiter=',';
input east north dat mmddyy8. elv elkid randid slope aspect GHF MAP sri LULC
SWE_100 Stdev_100 StudySWE StudyStdev ernd bf uf md;
Format date mmddyy8.;

If SWE_100<0 Then delete;
If Stdev_100<0 Then delete;

strt=1+ernd;

swea=SWE_100;
snha=Stdev_100;
sweasnha=swea*snha;
swel=StudySWE;
snhl=StudyStdev;

laswea=swea*swel;
lasnha=snha*swel;
lhswea=swea*snhl;
lhsnha=snha*snhl;

lhelv=snhl*elv;
lhsri=snhl*sri;
lhbf=bf*swel;lhuf=uf*swel;lhmd=md*swel;
```

```
snbf=bf*swea;snuf=uf*swea;snmd=md*swea;
sweasnha=swea*snha;
shbf=bf*snha;shuf=uf*snha;shmd=md*snha;
elvbf=bf*elv;elvuf=elv*uf;elvmd=elv*md;
sribf=bf*sri;sriuf=uf*sri;srimd=md*sri;

swea=((swea-0.43420)/0.434695);
snha=((snha-0.41252)/0.41252);
sweasnha=((sweasnha-0.13058)/.13058);
sri=((sri-339.647)/658.83);
elv=((elv-2331)/291);
lhelv=((lhelv-261.110)/261.231);
```

ADD MODELS

```
SWEA + (SWEA*SWEL)
SNHA + (SNHA*SWEL)
SWEA + SNHA + (SWEA*SWEL) + (SNHA*SWEL)
```

```
proc phreg data=pre;
strata dat;
model strt*ernd(1) = bf uf md/ ties=discrete;
title1 "model 1.1";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = sri/ ties=discrete;
title1 "model 1.2";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = elv/ ties=discrete;
title1 "model 1.3";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = bf uf md sri/ ties=discrete;
title1 "model 1.4";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = bf uf md elv/ ties=discrete;
title1 "model 1.5";

proc phreg data=pre nosummary;
```

```
strata dat;
model strt*ernd(1) = sri elv/ ties=discrete;
title1 "model 1.6";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = bf uf md sri elv/ ties=discrete;
title1 "model 1.7";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea/ ties=discrete;
title1 "model 2.1";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = snha/ ties=discrete;
title1 "model 2.2";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha/ ties=discrete;
title1 "model 2.3";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha sweasnha/ ties=discrete;
title1 "model 2.4";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea bf uf md/ ties=discrete;
title1 "model 3.1";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea bf uf md elv sri/ ties=discrete;
title1 "model 3.2";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md/ ties=discrete;
title1 "model 3.3";

proc phreg data=pre nosummary;
strata dat;
```

```
model strt*ernd(1) = swea snha bf uf md elv sri/ ties=discrete;
title1 "model 3.4";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea bf uf md snbf snuf snmd/ ties=discrete;
title1 "model 3.5";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea bf uf md lhbf lhuf lhmd/ ties=discrete;
title1 "model 3.6";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md snbf snuf snmd/ ties=discrete;
title1 "model 3.7";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md shbf shuf shmd/ ties=discrete;
title1 "model 3.8";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md lhbf lhuf lhmd/ ties=discrete;
title1 "model 3.9";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md sweasnha/ ties=discrete;
title1 "model 3.10";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md sweasnha snbf snuf snmd/
ties=discrete;
title1 "model 3.11";

proc phreg data=pre nosummary;
strata dat;
model strt*ernd(1) = swea snha bf uf md sweasnha shbf shuf shmd/
ties=discrete;
title1 "model 3.12";

proc phreg data=pre nosummary;
```

```
strata dat;
model strt*ernd(1) = swea snha bf uf md sweasnha lhbf lhuf lhmd/
ties=discrete;
title1 "model 3.13";

run;
```

Appendix B

AIC Values for Elk Models

	Prewolf	Col-Wolf	Est-Wolf
Landscape Models			
HBT	35364	12485	11995
SRI	36566	12792	12199
ELV	32663	11423	10315
HBT+SRI	35363	12481	11951
HBT+ELV	31713	11140	10012
SRI+ELV	32664	11417	10255
HBT+SRI+ELV	31711	11132	9957
Snow Models			
SWEA	32131	11466	10652
SNHA	35301	12424	12133
SWEA+SNHA	31756	11337	10576
SWEA+SNHA+(SWEA*SNHA)	31656	11267	10554
Landscape and Snow Models			
SWEA+HBT	32081	11446	10608
SWEA+HBT+ELV+SRI	30861	10856	9677
SWEA+SNHA+HBT	31736	11332	10543
SWEA+SNHA+HBT+ELV+SRI	30268	10664	9538
SWEA+HBT+(SWEA*HBT)	32020	11439	10596
SWEA+HBT+(SNHL*HBT)	32013	11402	10591
SWEA+SNHA+HBT+(SWEA*HBT)	31728	11334	10537
SWEA+SNHA+HBT+(SNHA*HBT)	31656	11304	10524
SWEA+SNHA+HBT+(SNHL*HBT)	31658	11290	10531
SWEA+SNHA+HBT+(SWEA*SNHA)	31628	11259	10516
SWEA+SNHA+HBT+(SWEA*SNHA)+(SWEA*HBT)	31628	11265	10514
SWEA+SNHA+HBT+(SWEA*SNHA)+(SNHA*HBT)	31562	11243	10503
SWEA+SNHA+HBT+(SWEA*SNHA)+(SNHL*HBT)	31573	11227	10505