**Data Reduction - Rocky Mountain & Artic Wolves**

This project followed a study completed by Pierre Jolicouer, called the ‘Multivariate Geographical Variation in the Wolf (Canis *lupus* L.) This study was specifically measuring morphometric data on Rocky Mountain and Artic wolves. Canis *lupus* or Gray Wolf has 40 sub species classified and is believed to be a species which shows a great deal of polymorphism geographically. Polymorphism qualities are potentially determined by three mechanisms; 1) Genetic Polymorphism, 2) Geographical conditioning, 3) Random order. While the Artic Wolf, Figure 1, is thought to be an ancestor of the Gray Wolf, native to the Canadian Artic, it is smaller in dimensions. Rocky Mountain wolves, Figure 2, are thought to be an ancestor to the wolves of Russia and Scandinavia which are traditionally a larger species of wolf. Figure 3 reveals comparisons between the European wolves vs. the North American characteristics found in the Canadian wolves.





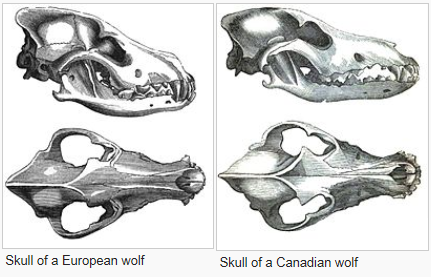


Figure 1

Figure 2

Figure 3

**Problem Statement:**

To develop a model based on Data Reduction techniques to find out what are the least set of morphometric variables needed to identify a particular species of wolves.

**Constraints and Limitations:**

The data collected was originally completed by Jolicoeur in 1958. The last since attempt to collect data on wolves was done in 1944. Limitations in the data come in the lack of use of multivariate processes to analyze geographic differentiation in the wolf. The data contained here is only pertained to morphological characteristics, the paper illustrate that “physiological, behavioral and ecological data could be done in similar manner”, but is missing in our analysis. Beyond possibly lacking potential variables, it “failed to show clearly nature and extent of geographical variation”. This is due to the data being purely observational and, as such, no casual inferences can be made about the relationship between the explanatory and response variable.

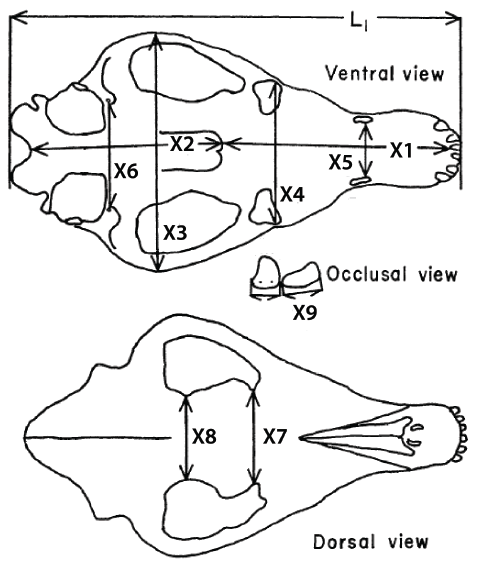
In addition, the data set contains only 25 observations of Artic and Rocky Mountain Wolves and which also is unbalanced in the two observations. The low sample might be due to various possible confounding factors, such as elusiveness of the animal or simply declining populations. As such we rely on Jolicoeur judgment to use morphological characteristics as significant variables of study.

A further limitation in this study, would be the lack of testing the sex variable fully in the project, I instead purely focus on statistical variability of wolf species.

**Data Set Descriptions:**

Our data’s response variable is ‘Location’ and it is broken into two types, Artic (AR) and Rocky Mountain (RM). To use that variable in our SAS program, we have coded into SAS a ‘Species’, Explanatory Variable (EV), AR = 1 / RM =2. Listed below are our EVs employed in this study, refer to Figure 4 for identification of measurement points employed:

Variables:



* Location: rm=rocky mountain / ar=arctic
* Sex: m=male f=female
* X1 = palatal length
* X2 = postpalatal length
* X3 = zygomatic width
* X4 = palatal width outside the first upper molars
* X5 = palatal width inside the second upper molars
* X6 = width between the postglenoid foramina
* X7 = interorbital width
* X8 = least width of the braincase
* X9 = crown length of the first upper molar
* Species = 1 for AR / 2 for RM

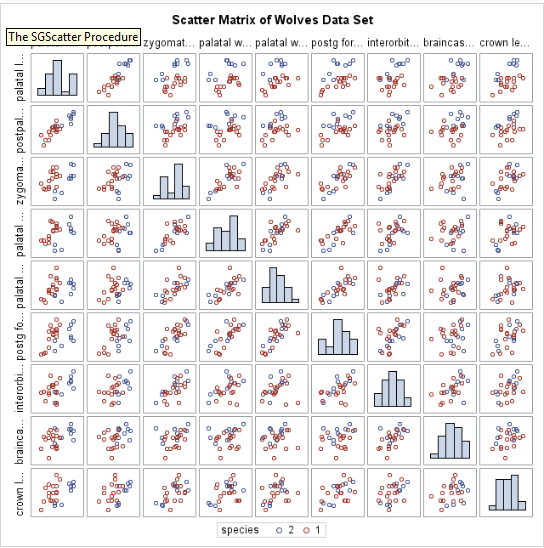
Figure 4

Each of these variables are common in either species morphological measurements and can impact certain possible trends in evolution of species, possibly being genetic, geographic, or random.

**Variable Screening (Correlation Matrix):**

We begin our analysis be examining the total structure of the data to see if an discrepancies exist which might need addressing before we begin data reduction steps.

Scanning the results from the correlation matrix in Fig. 5, we see normal distribution with strong linear trends from several EVs, such as X1 & X2, very strong and neatly. While the others are also showing positive trends and will need further examining, but all is useable and not in need of a transformation. Also our analysis of means (Fig. 6) also seems to be loaded correctly and but seems to differ in mean weights, especially X1 & X3.



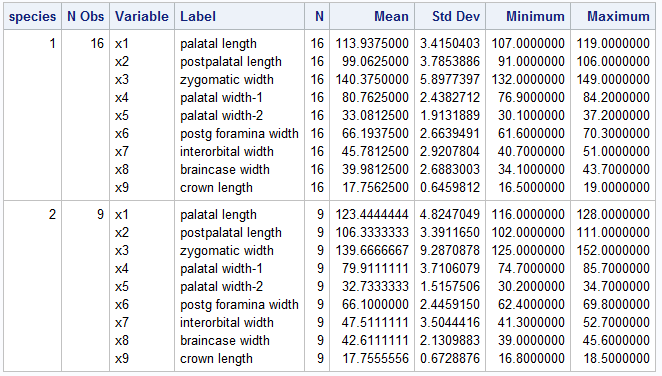


Figure 5

Figure 6

**Exploratory Data Analysis:**

For my process of selected an appropriate model I will attempt several selection methods. Principal Component Analysis (PCA), MANOVA, and the LASSO technique to see which model might best fit the data and also receive an appropriate Chi Square. So to start I will run PCA on the full model with all variables.

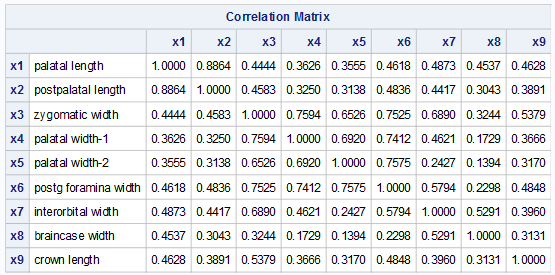
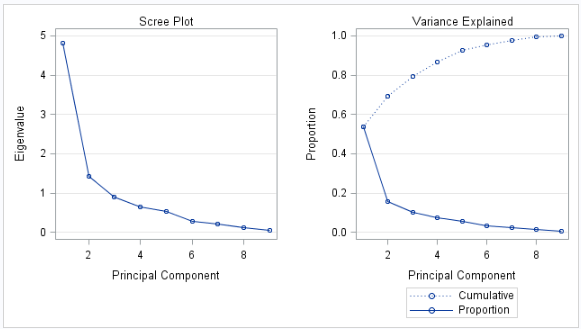


Figure 7

Running PCA, Fig 7, correlation Matrix show that the first load seems to be highly correlated with the first 2 EVs. And our Scree Plot in Fig. 8, also seems to demonstrate this assumption that the after the 2 component the line seems to level off, mirror what our Correlation Matrix states. This does however slight contradict the Eigenvector results, Fig.9, which seems to show that more than two EVs might be correlative. Studying PRIN1, we see X1, X2, X3, X4, X5, X6, X7 to all be quite close. Moving to PRIN2 it drops significantly, where X1 & X2 are all that seem to correlate.



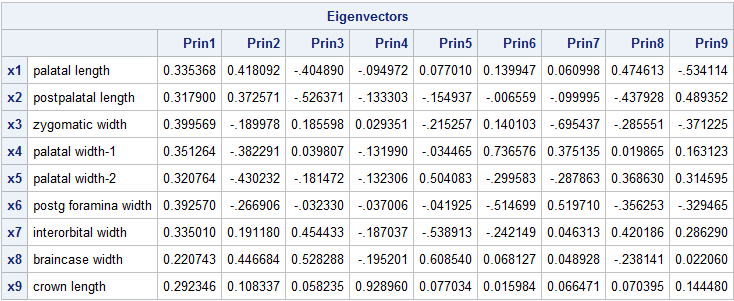


Figure 8

Figure 9

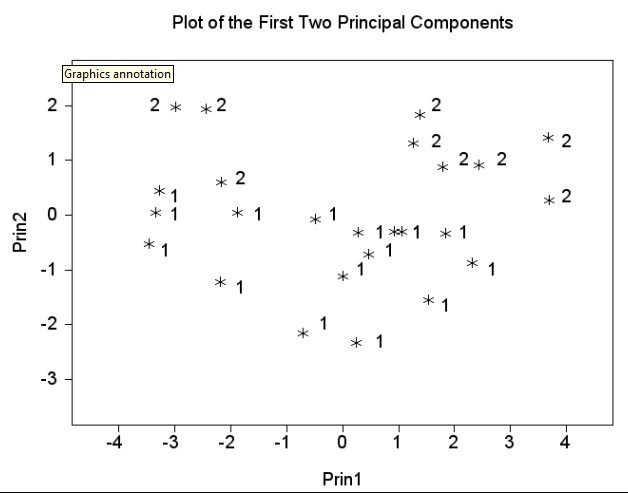
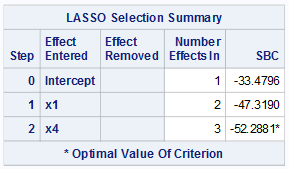


Figure 11



Figure 12



Studying the Plot of First two Principal Components, we can see that there are two distinct groups. Paired with a Full Model Chi Square test the full model does not fit within and we reject since we find a p-value of <.0001 significance. We will continue to push the model and find a Chi-Square which fails to reject.

Before we reduce, let us look at what LASSO selection gives. Refer to Fig. 12 we see that SAS is selecting X1, but surprisingly also X4, which never appeared significant previously.

Figure 10

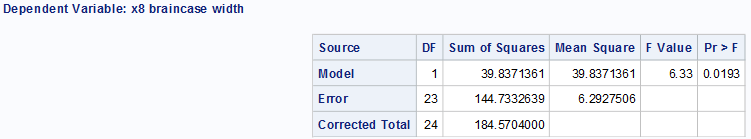
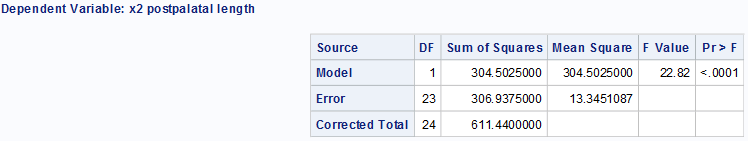
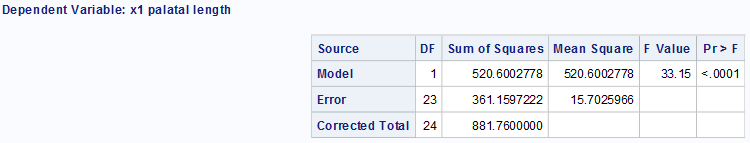
**Other METHODS**

In determining further steps, I choose to run MANOVA on the full model to see what would return significant. Not too much of a surprise as we saw that X1, X2, X8 all return significant as can be seen in Fig. 13-15.

Fig.15

Fig.14

Fig.13



**Model Selection:**

Now we will compare the 3 models which I have received as significant to see which ones have the best overall factors of significance dealing with p-values and R² and Adjusted R, additionally we will reexamine what Chi-Square are received as well.

The three models are shown below:

1. **Y*Species* = β₀ + β*X1(palatal length****)* **+ β*X2(postpalatal length)***
2. **Y*Species* = β₀ + β*X1(palatal length****)* **+ β*X4(postpalatal width)***
3. **Y*Species* = β₀ + β*X1(palatal length****)* **+ β*X2(postpalatal length)* + β*X8****(****least width of the braincase)***
4. **Y*Species* = β₀ + β*X1(palatal length****)* **+ β*X2(postpalatal length)* + β*X4****(postpalatal width****)***
5. **YSpecies = β₀ + βX1(palatal length) + βX2(postpalatal length)**



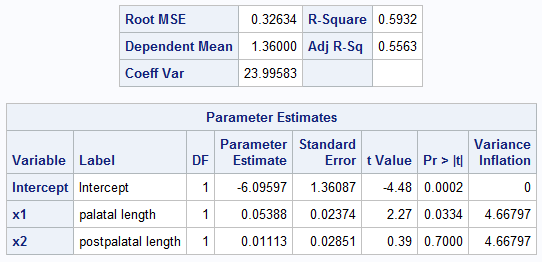
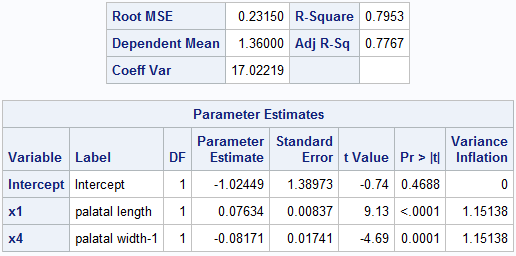


Figure 13

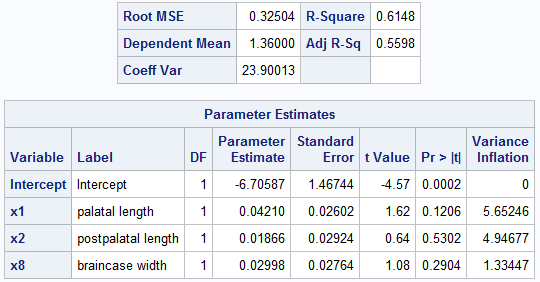
1. **YSpecies = β₀ + βX1(palatal length) + βX4(postpalatal width)**





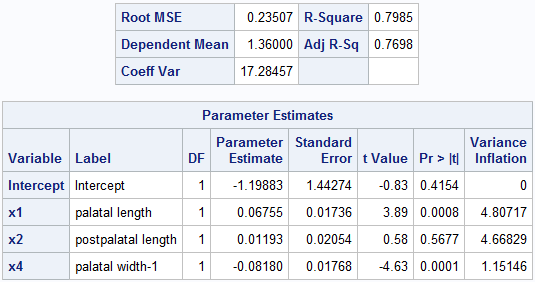
1. **YSpecies = β₀ + βX1(palatal length) + βX2(postpalatal length) + βX8(least width of the braincase)**





1. **YSpecies = β₀ + βX1(palatal length) + βX2(postpalatal length) + βX4(postpalatal width)**





After running the 4 different models, it seems that model #4 scores very well. In an effort to not to a “SAS Dump”, I will disclose that all model proved to be significant failing to prove the Null Hypothesis with a p-value <.0001. Model #4 was chosen because of the R² of 0.7985 and a significant 0.7002 provides strong evidence we can fail the Null Hypothesis of equal covariance matrices.

Going to the ‘Fit diagnostics’ (Figure 16) it is evident looking at our histogram and QQplot that the residual curve shows no signs against a normal distribution. The R-Student and Cook’s D only have one outlier, but we will keep since the group it is apart scores larger and is within the realm of possibilities.

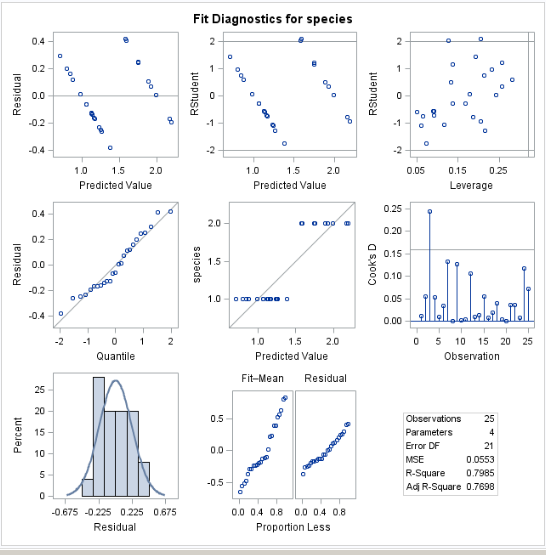


Figure 16

After examining for linearity, normality and constant variance our basic needs are reached. And find that this type of model which follows the width, length, depth of the skull, and width of brain cavity to be statistically significant variables with α= .05 level (n = 25, F = 29.06, p-Value = < .001).

**YSpecies = β₀ + βX1(palatal length) + βX2(postpalatal length) + βX4(postpalatal width)**

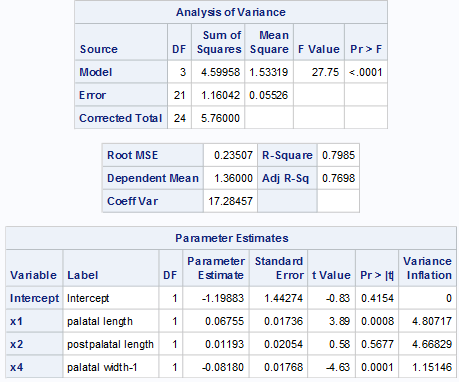


Figure 17

**Conclusion:**

This data set is my second time around and I have a different answer. I did find in my previous attempt that only the EV X1 was a suffiecinet point in determining a species. But this time, I am adding two more variables to the model, X2 & X4. Before I conclude, this final model has been

**DATA** wolves;

LENGTH location $**2** wolf $**5** sex $**1**;

INPUT location $ wolf $ sex $ x1-x9;

subject=\_n\_;

if location = 'ar' then species = **1**;

else if location = 'rm' then species = **2**;

LABEL

X1 = 'palatal length'

X2 = 'postpalatal length'

X3 = 'zygomatic width'

X4 = 'palatal width-1'

X5 = 'palatal width-2'

X6 = 'postg foramina width'

X7 = 'interorbital width'

X8 = 'braincase width'

X9 = 'crown length';

datalines;

rm rmm1 m 126 104 141 81.0 31.8 65.7 50.9 44.0 18.2

rm rmm2 m 128 111 151 80.4 33.8 69.8 52.7 43.2 18.5

rm rmm3 m 126 108 152 85.7 34.7 69.1 49.3 45.6 17.9

rm rmm4 m 125 109 141 83.1 34.0 68.0 48.2 43.8 18.4

rm rmm5 m 126 107 143 81.9 34.0 66.1 49.0 42.4 17.9

rm rmm6 m 128 110 143 80.6 33.0 65.0 46.4 40.2 18.2

rm rmf1 f 116 102 131 76.7 31.5 65.0 45.4 39.0 16.8

rm rmf2 f 120 103 130 75.1 30.2 63.8 44.4 41.1 16.9

rm rmf3 f 116 103 125 74.7 31.6 62.4 41.3 44.2 17.0

ar arm1 m 117 99 134 83.4 34.8 68.0 40.7 37.1 17.2

ar arm2 m 115 100 149 81.0 33.1 66.7 47.2 40.5 17.7

ar arm3 m 117 106 142 82.0 32.6 66.0 44.9 38.2 18.2

ar arm4 m 117 101 144 82.4 32.8 67.5 45.3 41.5 19.0

ar arm5 m 117 103 149 82.8 35.1 70.3 48.3 43.7 17.8

ar arm6 m 119 101 143 81.5 34.1 69.1 50.1 41.1 18.7

ar arm7 m 115 102 146 81.4 33.7 66.4 47.7 42.0 18.2

ar arm8 m 117 100 144 81.3 37.2 66.8 41.4 37.6 17.7

ar arm9 m 114 102 141 84.1 31.8 67.8 47.8 37.8 17.2

ar arm10 m 110 94 132 76.9 30.1 62.1 42.0 40.4 18.1

ar arf1 f 112 94 134 79.5 32.1 63.3 44.9 42.7 17.7

ar arf2 f 109 91 133 77.9 30.6 61.9 45.2 41.2 17.1

ar arf3 f 112 99 139 77.2 32.7 67.4 46.9 40.9 18.3

ar arf4 f 112 99 133 78.5 32.5 65.5 44.2 34.1 17.5

ar arf5 f 113 97 146 84.2 35.4 68.7 51.0 43.6 17.2

ar arf6 f 107 97 137 78.1 30.7 61.6 44.9 37.3 16.5

;

**RUN**;

**proc** **print** data = wolves;

**run**;

**proc** **means** data = wolves;

class species;

var x1 x2 x3 x4 x5 x6 x7 x8 x9;

**run**;

**proc** **univariate** data=wolves;

histogram;

**run**;

\*\*\*\*\*Scatterplot Matrix Testing Variables\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

;

title "Scatter Matrix of Wolves Data Set";

**proc** **sgscatter** data = wolves;

matrix x1 x2 x3 x4 x5 x6 x7 x8 x9 / diagonal=(histogram) group=species;

**run**;

ods listing style = rtf;

ods graphics on;

**proc** **princomp** data = wolves plots out=pcwolves;

var x1 x2 x3 x4 x5 x6 x7 x8 x9;

title 'Plot of the First Two Principal Components';

%***plotit***(data=pcwolves, labelvar=species,

plotvars=Prin2 Prin1, Color=black, color=blue);

**run**;

**proc** **discrim** data = wolves outstat=wolvestat

wcov pcov method=normal pool=test

manova listerr crosslisterr;

class species;

var x1 x2 x3 x4 x5 x6 x7 x8 x9;

**run**;

**proc** **reg** data=wolves;

title 'Reg Analysis / AIC VIF CLI';

model species = x1 x2 x3 x4 x5 x6 x7 x8 x9 / AIC VIF CLI;

**run**;

\*\*\*\*\*\*

MANOVA;

**proc** **glm** data =wolves;

class species;

model x1 x2 x3 x4 x5 x6 x7 x8 x9 = species;

manova h = species;

**run**;

**proc** **glmselect** data = wolves;

model species = x1 x2 x3 x4 x5 x6 x7 x8 x9 / selection = LASSO;

**run**;

\*\*\*\*\*\*\*\*Other methods\*\*\*;

**proc** **discrim** data = wolves pool=test crossvalidate;

class species;

var x1 x2 x3 x4 x5 x6 x7 x8 x9;

**run**;

**proc** **reg** data=wolves;

title 'Reg Analysis / AIC VIF CLI';

model species = x1 x2 / AIC VIF CLI;

**run**;

**proc** **discrim** data = wolves outstat=wolvestat

wcov pcov method=normal pool=test

manova listerr crosslisterr;

class species;

var x1 x2;

**run**;

**proc** **reg** data=wolves;

title 'Reg Analysis / AIC VIF CLI';

model species = x1 x4 / AIC VIF CLI;

**run**;

**proc** **discrim** data = wolves outstat=wolvestat

wcov pcov method=normal pool=test

manova listerr crosslisterr;

class species;

var x1 x4;

**run**;

**proc** **reg** data=wolves;

title 'Reg Analysis / AIC VIF CLI';

model species = x1 x2 x8 / AIC VIF CLI;

**run**;

**proc** **discrim** data = wolves outstat=wolvestat

wcov pcov method=normal pool=test

manova listerr crosslisterr;

class species;

var x1 x2 x8;

**run**;

**proc** **reg** data=wolves;

title 'Reg Analysis / AIC VIF CLI';

model species = x1 x2 x4 / AIC VIF CLI;

**run**;

**proc** **discrim** data = wolves outstat=wolvestat

wcov pcov method=normal pool=test

manova listerr crosslisterr;

class species;

var x1 x2 x4;

**run**;

\*Leave out CV;

**proc** **glmselect** data = wolves;

model species = x1 x2 x4 / selection = forward(STOP=Press);

**run**;

\*10 Fold CV;

**proc** **glmselect** data = wolves;

model species = x1 x2 x4 / selection = forward(Choose=CV)CVmethod=Random(**10**);

**run**;

\*5 Fold CV;

**proc** **glmselect** data = wolves;

model species = x1 x2 x4 / selection = forward(STOP=Press);

**run**;

**Bibliography:**

1. Subspecies of Canis lupus. In *Wikipedia, the free encyclopedia*,

2/14/2016 12:09 PM from <https://en.wikipedia.org/wiki/Subspecies_of_Canis_lupus>

1. Jolicoeur, Pierre (1959). *Multivariate Geographical Variation in the Wolf Canis lupus L.* Vancouver: Society for the Study of Evolution.