Stephen Mayor

Snow hardness in response to scale:

Using the Generalized Linear Model to compare means of four groups

Introduction

Analysis of Variance (ANOVA) is a standard, commonly practiced statistical procedure. It is a special case of the General Linear Model (GLM), and therefore assumes homogeneous residuals, a normal error structure, and an identity link. If these assumptions are invalid, the GLM cannot be applied. However, if the residuals are distributed in way that fits a common distribution pattern, such as gamma, poisson, or binomial, then the Generalized Linear Model (GzLM) can be applied by specifying this distribution. Using the GzLM, an analysis of deviance (ANODEV) can be employed, which replaces the ANOVA.

In this analysis, I was interested in how snow harness varied in response to the group of measurements it was taken from. The groups of measurements were four different behavioural scales of selection by woodland caribou of the Middle Ridge herd in the Bay du Nord area of southern Newfoundland. The scales of selection were: craters (where caribou have dug into the snow to access food), feeding sites (uncratered sites in areas of abundant craters), travel routes (sites where caribou have traveled between feeding areas), and winter range (sites systematically sampled within the area used by caribou). I was interested in whether the relationship between each of these scales and snow hardness was significant, and whether the snow hardness was significantly different between each of these scales.

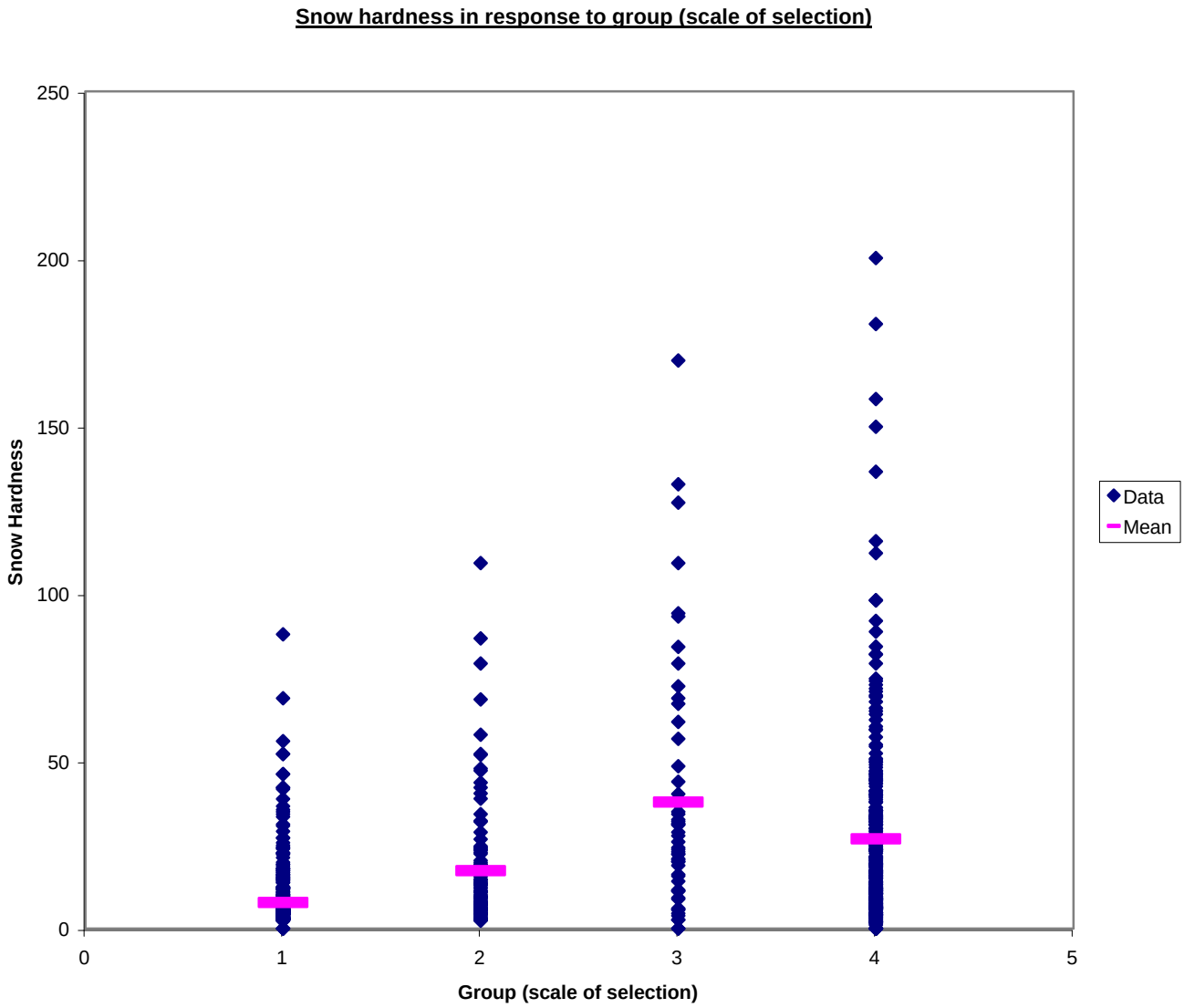
1. Construct model

Response variable: Snow hardness (SH) is a ratio scale variable and was measured in g/cm^2 using a ram penetrometer.
Explanatory variable: Group (Grp) is a unitless, nominal scale variable.
Four groups, each with a mean, are used in this analysis. They are: (1) Craters, (2) Feeding sites, (3) Travel routes, (4) Winter Range and represent different scales of selection by caribou.

Verbal model:
Snow hardness depends on group (scale of selection).

Formal model:
$SH = ß_O + ß_{X\,Grp}\, X_{Grp} + res$

Graphical model: comparison of means in four groups



The error structure is assumed to be normal.

2. Execute model

The general linear model was used, which specifies a normally distributed error structure and an identity link function.

Using SAS, the model was executed using PROC GENMOD.

```
proc genmod;
class Group;
model SnowHard= Group/
dist=normal
link=identity
type1
type3 obstats residuals;
ods output obstats=resids;
```

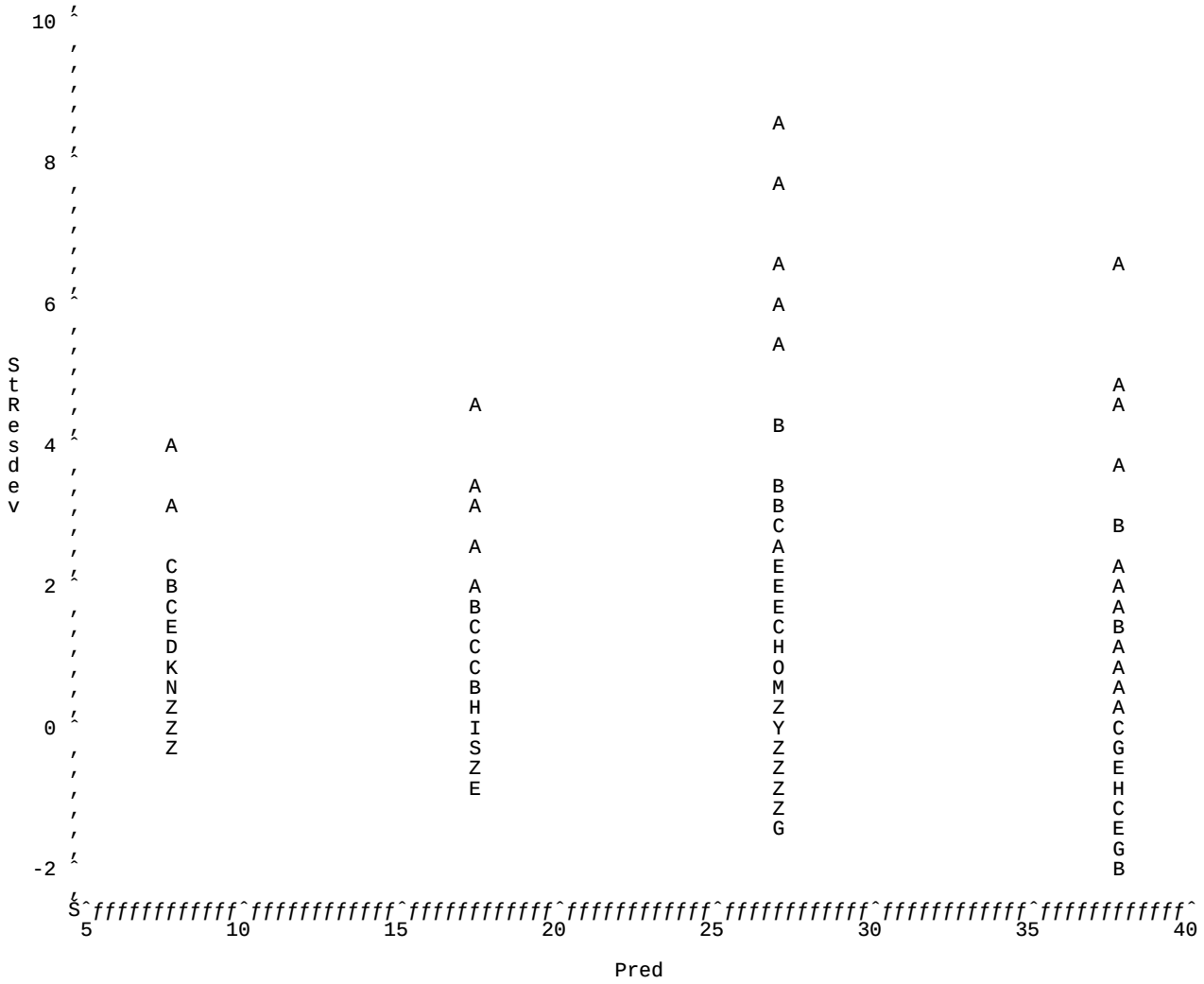
I used a Type III analysis, because order of variables is not important and is often more appropriate for field experiments. Type I analyses, by contrast, depends on the order of explanatory variables.

3. Evaluate the model

Homogeneity assumption:
Figure 1, the residuals vs. fits plot, shows a moderate cone shape in the data, indicating that the errors may not be homogeneous.

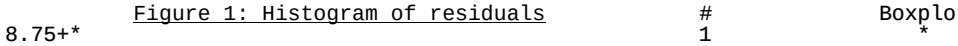
Straight line assumption:
Figure 1 does not show evidence of bowls or arches, indicating that a straight line model is appropriate.

Figure 1: Plot of Stresdev*Pred. Legend: A = 1 obs, B = 2 obs, etc.



Normality assumption:
Figure 1 and Figure 2 show that the residuals are not normal but highly skewed.

Figure 1: Histogram of residuals



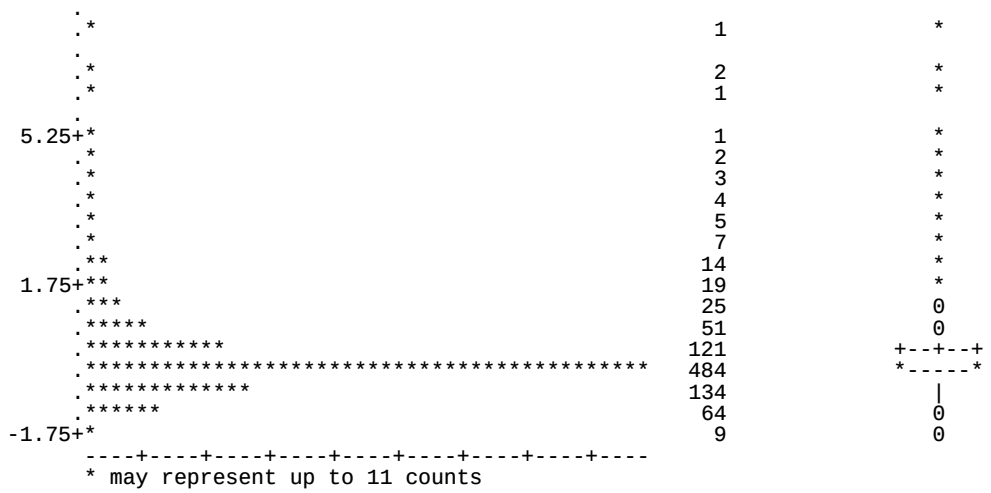
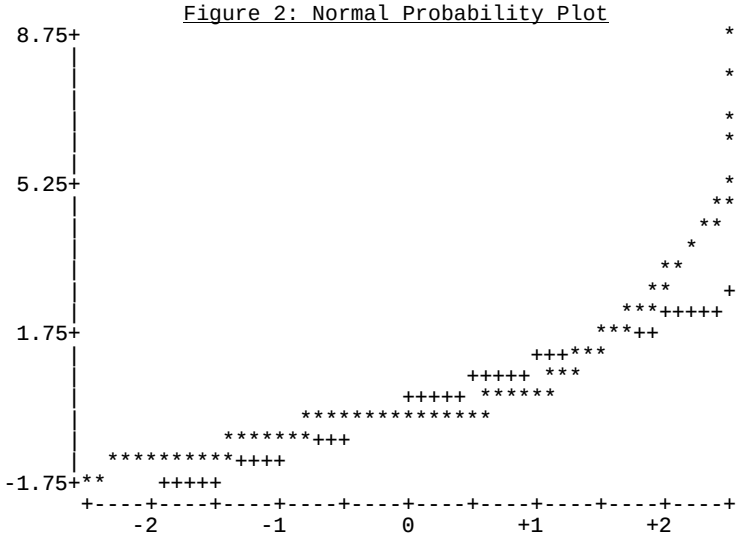


Figure 2: Normal Probability Plot



Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	26.8023	1.1606	24.5275	29.0771	533.28	<.0001
Group 1	1	-19.0583	1.4714	-21.9422	-16.1743	167.76	<.0001
Group 2	1	-9.5534	2.3300	-14.1202	-4.9866	16.81	<.0001
Group 3	1	10.9626	2.9487	5.1833	16.7420	13.82	0.0002
Group 4	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale	1	20.1028	0.4617	19.2180	21.0283		

NOTE: The scale parameter was estimated by maximum likelihood.

Because the assumptions of homogeneity and normality are invalid, the model must be revised. The general linear model cannot be used, so instead I use the more flexible generalized linear model, and re-start the ten-step recipe at step 1.

1. Construct model

The model is the same as before, however I revised the error structure.
Gamma is appropriate for a continuous but skewed distribution.

The identity link was used again. The identity link is appropriate because there is additive change in the variables. That is, each variable changes by an increment, not multiplicatively.
The link function in generalized linear models specifies a nonlinear transformation of the predicted values so that predicted values fit the specified distribution, in this case gamma. The link function is therefore used to model responses when a dependent variable is assumed to be nonlinearly related to the predictors. The link function serves to link the random component of the model, the probability distribution of the response variable, to the systematic component of the model (the linear predictor).

2. Execute model

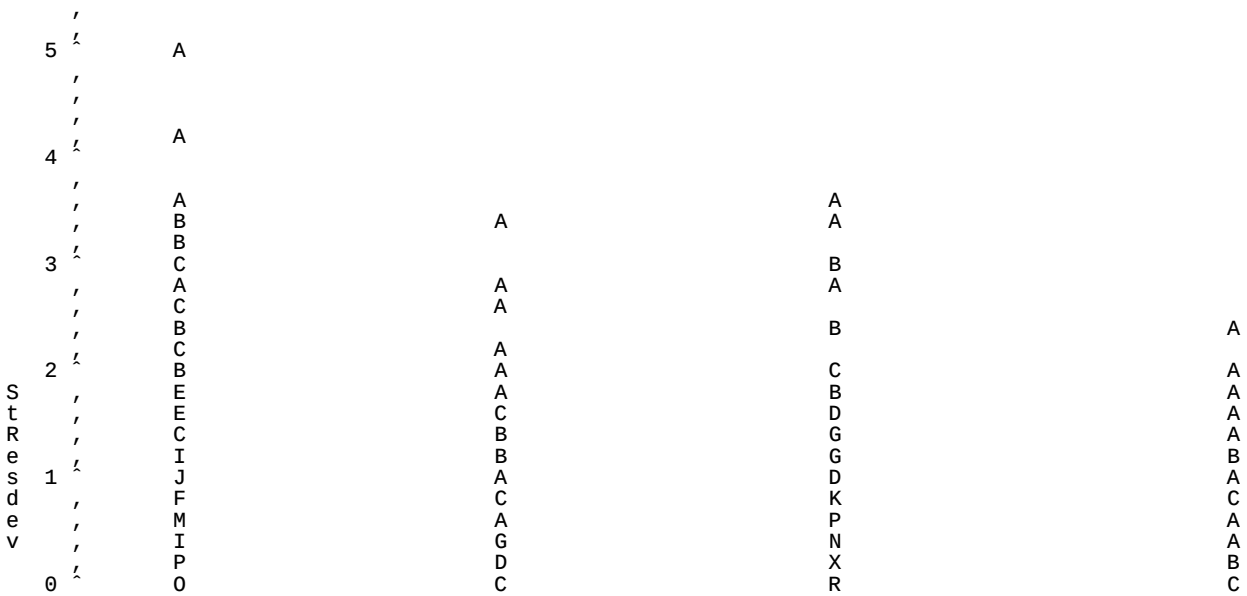
The gamma distribution and identity link were used to implement the generalized linear model in SAS using Proc genmod.

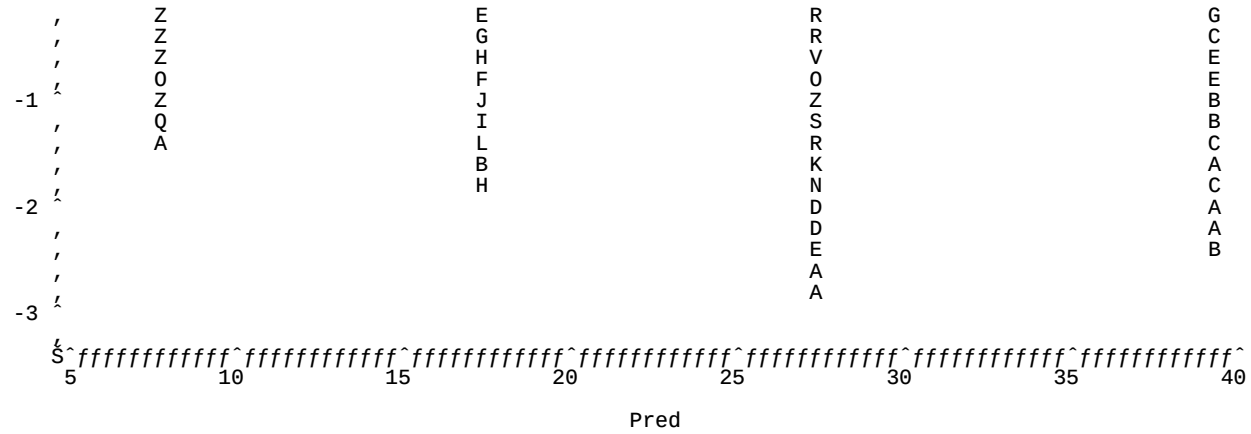
```
proc genmod;
class Group;
model SnowHard= Group/
dist=normal
link=identity
type1
type3 obstats residuals;
ods output obstats=resids;
```

3. Evaluate model

Figure 4 shows the residuals vs fits plot of the data. Although there appears to be a negative relationship, there are no bowls or arches, indicating a valid straight line assumption, and no cones, indicating a more homogeneous distribution

Figure 4: Plot of Stresdev*Pred. Legend: A = 1 obs, B = 2 obs, etc.





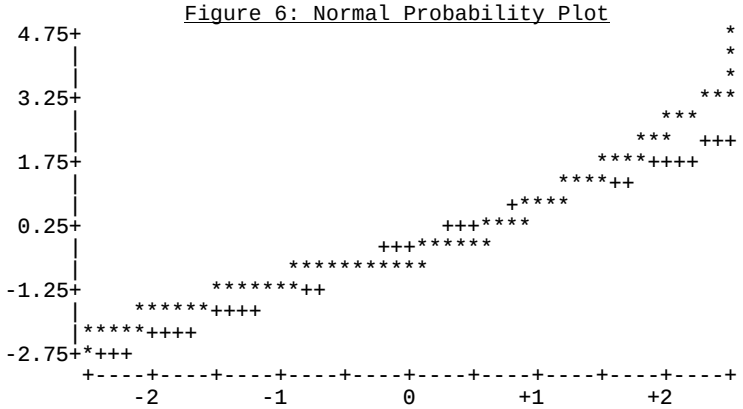
NOTE: 12 obs had missing values. 242 obs hidden.

Figures 5 and 6 show that the residuals are nearly normal. Although the residuals are still slightly skewed, the fit of the gamma distribution is a big improvement over the normal distribution.

Figure 5: Histogram of residuals

```
                                                #             Boxplot
4.75+*                                          1                *
    .*                                          1                *
    .*                                          2                *
3.25+*                                          7                *
    .**                                        11                0
    .**                                        11                0
1.75+****                                      27                0
    .*****                                     40                |
    .********                                  63                |
0.25+************                              90           +-----+
    .*********************                    164          |  +  |
    .****************************************  343          *-----*
-1.25+***************                         118               |
    .******                                    42                |
    .**                                        14                0
-2.75+*                                         2                0
     ----+----+----+----+----+----+----+----+---
     * may represent up to 8 counts
```

Figure 6: Normal Probability Plot



4. State sample and population

The population is the total number of snow hardness measurements that could have been taken in the Middle Ridge area during the winter of 2005 at various scales of selection. The sample is thought to be representative.

5. Hypothesis testing as mode of inference.

Hypothesis testing is appropriate because I am interested in the variance between groups.

6. State H_A / H_o pairs for model

H_A: $Var(ß_{XGrp}) > 0$ H_O: $Var(ß_{XGrp}) = 0$

Tolerance for Type I error, a = 0.05

7. ANODEV

The Analysis of Deviance table summarizes information about the sources of variation in the response for the set of data. The df were partitioned according to the model.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Group	3	464.75	<.0001

The change in goodness of fit from the intercept is 464.75, which is statistically significant because p<0.0001 and p < a.

8. Recompute p if necessary

Recomputation of p is not necessary because assumptions are reasonably met, the sample is large, and p is not near a.

9. Declare decision about model terms

Accept H_A: $Var(ß_{XGrp}) > 0$ and reject H_O: $Var(ß_{XGrp}) = 0$ because p< a.

10. Examine parameters of biological interest

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	27.4426	1.2761	24.9416	29.9437	462.48	<.0001
Group	1	1	-19.6513	1.3064	-22.2118	-17.0907	226.26	<.0001
Group	2	1	-10.1937	1.8795	-13.8774	-6.5100	29.42	<.0001
Group	3	1	11.7473	4.4707	2.9849	20.5098	6.90	0.0086
Group	4	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	1.5784	0.0667	1.4530	1.7147		

NOTE: The scale parameter was estimated by maximum likelihood.

Each group (scale of selection) was statistically significant.
The mean of each group is significantly different from all but one other group, with uncertainty measured by confidence limits. Therefore the snow hardness observed in craters and feeding sites is significantly different from that in the winter range. Therefore caribou are selecting for lower snow hardness.

Group	**Snow hardness (mean)**	**Wald 95% Confidence Limits**	
Craters	7.7 a b x	2.7298	12.853
Feeding Sites	17.2 c d x	11.0642	23.4337

Travel Routes	39.1 a c x	27.9265	50.4535
Winter Range	27.4 b d x	24.9416	29.9437

x = significant p

a, b, c = estimates with same subscript are significantly different from each other

Note: Use of GzLM with multivariate data

The ten step process was repeated for other biologically important variables related to caribou habitat selection, including the first principle component extracted from a PCA, which explained 12.5% of the variance in the data. Principle components analysis (PCA) reduces multivariate data to a smaller number of underlying variables, using correlations between variables to eliminate redundancy. A summary of the means and their significance follows. The table shows that significance at various levels can be compared across variables, even when parameters were assessed with different distributions.

Group	**Snow hardness**	**Snow depth**	**Cladina**	**Kalmia**	**PCA1**
Craters	7.7 a b x	14.4 a b x	68.2 a x	23.1 a b x	9.5 a b x
Feeding Sites	17.2 c d x	20.7 c x	58.1 b x	27.8 a x	9.7 c d x
Travel Routes	39.1 a c x	31.7 a x	24.2 a b c x	19.5 x	10.8 a c x
Winter Range	27.4 b d x	28.2 b c	25.2 a b c	17.7 b	10.8 b d
Distribution, link	*Gamma, identity*	*Gamma, identity*	*Normal, identity*	*Gamma, identity*	*Gamma, identity*

x = significant p

a, b, c = estimates with same subscript are significantly different from each other