1998

# Validating a Health Questionnaire for Predicting Neuropathy in Patients with Insulin-Dependent Diabetes Mellitus

Matthew J. Maurer
*College of Saint Benedict/Saint John's University*

# VALIDATING A HEALTH QUESTIONNAIRE FOR PREDICTING NEUROPATHY IN PATIENTS WITH INSULIN-DEPENDENT DIABETES MELLITUS

**A THESIS**

**The Honors Program**

**St. John's University**

**In Partial Fulfillment of the Requirements for the Distinction of "All College Honors" and**

**the Degree Bachelor of Arts**

**In the Department of Mathematics**

**By**

**Matt Maurer**

**April 17, 1998**

PROJECT TITLE: Validating a Health Questionnaire

Approved by:

1st line - Signature of your project advisor

Robert Dumoneaux

Regents Professor of Mathematics

Philip J. Ry

Professor of Mathematics

Mario A Bado

Assistant Professor of Mathematics

Philip J Ry

Chair, Department of Mathematics

Margaret L Cook

Director, Honors Thesis Program

Charles A. Bolduc

Director, Honors Program

# TABLE OF CONTENTS



## TABLE OF EQUATIONS



## TABLE OF FIGURES



## ACKNOWLEDGEMENTS


I am very grateful to Professor Peter Cavanagh and Dr. Robert Van Deursen of the Center for Locomotion Studies (CELOS) at Penn State University for permitting me to use the data and questionnaire from their Posture in Diabetics Study. Thanks are also due to Mary Becker for information regarding the CELOS diabetes study and introduction to the workings of CELOS. I would especially like to thank John O'Gorman, Dr. Janice Derr, and Dr. James Rosenberger in the Penn State Department of Statistics for their assistance, guidance, and giving me the opportunity to do my own research.

## 1. Abstract

Questionnaires are a cost-effective method for screening large numbers of people for health problems. More expensive clinical follow-up can focus on people whose responses to the questionnaire suggest they are most at risk. To my knowledge, no questionnaire has ever been developed to screen for neuropathy in diabetics. Using the questionnaire developed by Dr. Peter Cavanagh and Dr. Robert Van Deursen at the Center for Locomotion Studies (CELOS) at Penn State University, I was able to create a model from the questionnaire that predicts the presence of neuropathy. The model has a sensitivity of 92.9%, correctly diagnosing nearly 93% of the neuropathics in the study. Its overall accuracy is 72.7%, with an ROC AUC of .813. Reliability, however, is somewhat low with a kappa value of .339. The model is based on two questions: "Do you currently have a decrease in the strength of your legs or feet that is out of proportion with any general changes in your overall strength?" and "Do you have numbness in your feet?" Diabetics having decreases in strength were 10.4 times more likely to have neuropathy. Diabetics having numbness in their feet were 4.58 times more likely to have neuropathy. Also, for a diabetic that has both a decrease in strength and numbness, the probability of having neuropathy is about .92.

I also examined the questionnaire for possible changes. I reduced the number of levels of several questions, as many levels were not used by any of the participants in the study. These questions may be changed on the questionnaire to reflect the question level adjustment. I feel

this questionnaire may be used as a preliminary device for screening diabetics for neuropathy. Also, I expect improvements in the model when more data becomes available.

## 2. Introduction

A major problem of diabetics is nerve death, also known as neuropathy. The decreased function in nerves causes loss of feeling and sensation. This is especially prevalent in the hands and feet, as the longest nerves are the first to be affected. The loss of sensation in the feet is of particular interest in this study. Lack of feeling in the feet can lead to problems with walking and balance, which can cause falls or injuries. Other, more serious conditions can arise; in some cases amputation is necessary. These problems can lead to a change in lifestyle and a decrease in general level of health for the diabetic.

A major issue with neuropathy is the loss of protective sensation. People with limited sensation in their feet may walk all day with a rock in their shoe, or a fold in their sock. While unaffected people would stop to take the rock out, or adjust their gait pattern to adjust to the sock fold, neuropathics are oblivious to the problem and will not notice the irritant until the end of the day when they take their shoe off. These irritants can cause swelling, blisters, or more serious ulcers and tissue damage. People with neuropathy can prevent problems by taking special care of their feet and checking their shoes and socks often for debris and folds.

Due to the health risks associated with unattended neuropathy, it is advantageous to diagnose diabetics having neuropathy. The risk of neuropathy increases as the time of having diabetes increases. Pirat (1979) has shown that up to 50% of all diabetics who have had diabetes

for at least twenty years and Orchard et al. (1990) has shown up to 70% of those with diabetes for thirty years have some level of neuropathy. Thus, neuropathy is a very common health concern among long term diabetics. There exist standard, documented clinical tests for neuropathy. However, these tests require a trained administrator with special tools. Not all people have the time or resources to be clinically tested on a regular basis. Thus, a preliminary screening questionnaire to predict neuropathy presence would be a valuable tool for diabetics at risk for neuropathy.

Questionnaires are very common and useful in the health science field. They are inexpensive, easy to use, and can be administered to large numbers of people. Most are not meant to supplant standard clinical measures. Rather, they serve as preliminary screening devices. Those scoring positive on the screening test may be referred for the standard clinical evaluation. There exist several comprehensive guides to health questionnaires. The references I found most useful in my research are Streiner and Norman (1989), a statistical guide to developing a health questionnaire, and McDowell and Newell (1996), a volume full of documentation with a large assortment of health and social science questionnaires with commentary on each scale.

## 3. Questionnaire

The questionnaire for the study was developed by Dr. Peter Cavanagh, Dr. Robert Van Deursen, and Mary Becker at the Center for Locomotion Studies (CELOS) at Penn State University, during a three year study of diabetics, their postures and falls. The questionnaire has been used in other studies, most recently Van Deursen (1997). In Van Deursen, though, the

questionnaire was not used in the role I used in my study. The questionnaire consists of 32 questions, roughly grouped in eight sections. Questions deal with activity level, falls, balance, comfort and ease of aspects of daily living, and feeling in legs. Questions range from 2 levels to 5 levels, with questions within each section usually having the same number of levels. A copy of the questionnaire can be found in Appendix 9.2.

## 4. Subjects

Subjects for the questionnaire survey were taken from diabetics from Pennsylvania. Initially, 182 applicants responded to calls for subjects. Potential subjects were interviewed by phone. Subjects with major health problems: strokes, recent surgeries, etc. were eliminated. Potential subjects were also screened for age, weight, medication, and other exclusion factors. Qualified subjects were brought to CELOS for a clinical examination. Of the 182 applicants, 46 diabetics were used in the study. Subjects were clinically evaluated and classified into three groups: non-neuropathic, mild neuropathic, and severe neuropathic. The neuropathy classifications are based on BVPT values. This is a standard procedure for identifying neuropathy, and has been well documented. A control group of 15 non-diabetic, non-neuropathic subjects was also part of the study, creating a total of four groups. A breakdown of the groups can be found in Figure 1.

**Figure 1 Group Description**

| Group | N | Gender M/F in % | Age (years) | Height (m) | Weight (kg) |
|---|---|---|---|---|---|
| 1 ND | 15 | 66.7/33.3 | 59.13 <53.61, 64.66> | 1.717 <1.672, 1.763> | 81.9 <77.9, 85.8> |
| 2 NNP | 16 | 58.8/41.2 | 59.53 <54.59, 64.47> | 1.675 <1.614, 1.736> | 81.4 <72.1, 90.8> |
| 3 MNP | 14 | 78.6/21.4 | 57.21 <50.14, 64.28> | 1.735 <1.692, 1.778> | 88.3 <74.9, 101.7> |
| 4 SNP | 15 | 80.0/20.0 | 64.67 <57.14, 66.19> | 1.751 <1.702, 1.799> | 89.7 <82.6, 96.8> |

**Subject Characteristics: Percentage of each gender, means, and 95% confidence intervals for age, height, and weight for each group.**

## 5. Methods

Since the questionnaire was to be administered to diabetics, I deleted the non-diabetics (group 1) from our database. Though this significantly reduced the sample size, it is a necessary step. Including the non-diabetics would produce an inaccurate model for diabetics, as the questionnaire would probably not be given to non-diabetics. I then collapsed the remaining three groups into two, non-neuropathic, and neuropathic. In future studies, with larger data, it may be possible to create a model to classify neuropathy level as well as identify neuropathic and non-neuropathic subjects. However, that is beyond the scope of the present study.

This resulted in a new sample size of 45, with 29 neuropathic and 16 non-neuropathic subjects. Two subjects did not fill out a questionnaire; both were neuropathic, originally group 3 (mild neuropathic). These subjects were dropped from the data set, leaving a sample size of 44, with 28 neuropathic.

Descriptive statistics and cross tabulations were computed for each individual question, (Appendix 9.3). In several questions, not every level was used. In these cases, I collapsed the levels of the question. The questions were analyzed by cross tabulations using the chi-squared test. Due to the small sample size, many questions had predicted counts of less than 5, invalidating the chi-squared test. In these cases, Fisher's exact test was used.

Questions with $p>.25$ from the $X^2$ test or Fisher's exact test were initially eliminated. The remaining questions were placed into a logistic regression program to create a model for predicting neuropathy. Possible models were generated and examined. The model's validity was measured using receiver operating characteristic curves, and the model's reliability was measured using the kappa computation for reliability. A discussion on the model tool, validity, and reliability follows.

## 5.1. Logistic Regression

Logistic regression is powerful tool to analyze binary categorical data, when the response is either a 'success' or 'failure'. Since my data is classified in two categories, non-neuropathic and neuropathic, binary logistic regression is a logical tool for creating a model to predict the presence of neuropathy. I am including a brief description of the logistic regression model. A

more in-depth treatment can be found in Agresti (1990), Agresti (1996) or Hosmer and Lemeshow (1989). The Hosmer and Lemeshow reference is a very explanatory, comprehensive covering of logistic regression.

In logistic regression, variables are used to create a model to describe the relationship between a binary response variable and predictor variables. I will use a univariate example for explanation, but the model can contain, and usually does, several predictor variables. The expected value of the response variable, Y, at a given predictor variable value, x, is denoted as E(Y|x). In logistic regression, the expected value, or mean, has domain of [0,1]. For purposes of simplification, in logistic regression, the expected value is labeled $\pi(x)$. This is what we end up estimating.

**Equation 1 Expected value of Y, given x**

$$y_i = \pi(x) + \varepsilon_i$$

Note: the error term, $\varepsilon$, is distributed with mean zero and variance $\pi(x)[1-\pi(x)]$. Since $\pi(x)$ has domain [0,1], we can use the logistic curve to model $\pi(x)$.

**Equation 2 Logistic curve**

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

As in linear regression, it is useful to transform the data to yield a linear model. A transformation of $\pi(x)$, the logit, is used. This transformation is defined by

**Equation 3 The logit**

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = ß_0 + ß_1 x$$

Note that the transformation has the reals for a domain. From this transformation, we have $\pi(x)$ transformed so that it is linear in terms of the x's. The coefficient of the predictive variable, $ß_1$, represents $\ln(\psi)$, where $\psi$ is the odds ratio:

**Equation 4 The odds ratio**

$$\Psi = e^{g(x)} = \frac{\pi(x)}{1-\pi(x)}$$

and $\ln(\psi)$ is the log-odds. The odds ratio intuitively represents how much more likely a success will occur when the value of the prediction variable increases by one unit. Consider a univariate model, using a single question for our study, in which $ß_1$=1.45. A subject answering with a response coded as 1 is exp(1.45) = 4.26 more likely to have neuropathy than a subject answering with a response coded as zero. So, as a coefficient approaches zero, for a variable having no effect on the response, the odds ratio should approach exp(0)= 1. This makes sense, since a change in an unaffecting variable will not affect the odds of the outcome.

Maximum likelihood theory is used to estimate the parameters $ß_0$ and $ß_1$. The derivation of $ß_0$ and $ß_1$ is iterative in nature and is computed by most standard statistical packages, including Minitab and SAS, the packages I used in my work. The maximum likelihood estimates are those values of $ß_0$ and $ß_1$ that maximize the probability or likelihood of observing the results

of the study. These estimates are labeled $\hat{ß}_0$ and $\hat{ß}_1$. An estimate, $\hat{p}$, of the probability the response variable is 1 or a success from the prediction variables can be made using the coefficient estimates.

**Equation 5 Estimated probability of Y given x**

$$\hat{p} = \hat{\pi}(x) = \frac{e^{\hat{ß}_0 + \hat{ß}_1 x}}{1 + e^{\hat{ß}_0 + \hat{ß}_1 x}}$$

This value can be used to assess likelihood of the response occurring, and can be used as cutoff values for scales.

### 5.2. Validity and Reliability

Validity and reliability are two important concepts when dealing with health questionnaires. Validity is a measurement of the overall accuracy of the model. A model needs to correctly predict the presence or absence of a condition to be useful, thus a high level of validity is required. A test must also assess subjects consistently. If a subject is given the questionnaire on multiple occasions, the results should be highly correlated. This is the concept of reliability. A test must be reliable to be effective. For a more in-depth discussion on validity and reliability, see Streiner and Norman (1989) or McDowell and Newell (1996).

Two important concepts of validity and diagnostic utility are sensitivity and specificity. The values of sensitivity and specificity are key assessments of the ability of the model in prediction of the response variable. Sensitivity is how accurately the model predicts those with the response variable. Numerically, it is

### Equation 6 Sensitivity

$$\frac{\text{Number of events predicted as events (true positives)}}{\text{Total number of events}}$$

In our model, it is the percentage of neuropathics that the model labels neuropathic. These are called true positives. Specificity, on the other hand, measures how discriminating the model is. Numerically, it is

### Equation 7 Specificity

$$\frac{\textit{Number of failures predicted correctly (true negatives)}}{\textit{Total number of failures}}$$

In our model, it is the percentage of non-neuropathics that the model labels non-neuropathic. The quantity (1-Specificity) is commonly used in plots to analyze the effectiveness of the model. This quantity is

### Equation 8 1-Specificity

$$\frac{\textit{Number of failures predicted falsely (false positives)}}{\textit{Total number of failures}}$$

In our model, it is the percentage of non-neuropathics that are labeled as neuropathic.

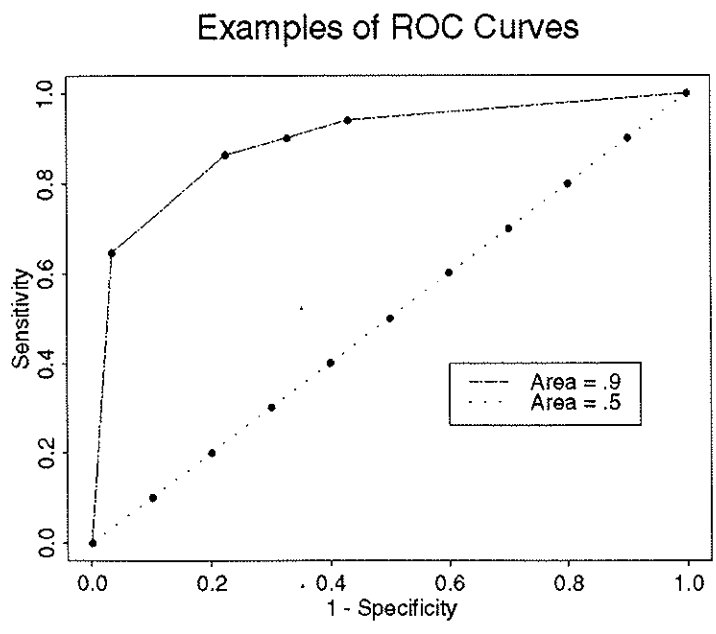
The sensitivity and specificity vary with the $\hat{p}$ from the logistic regression model. A low $\hat{p}$ cutoff from the model will generally have high sensitivity and low specificity, since most test subjects will have a $\hat{p}$ above the cutoff value. Likewise, a high $\hat{p}$ cutoff will generally have a

high specificity and low sensitivity. It is desirable to have a scale that is high in sensitivity and specificity, as the overall accuracy of the test is dependent on both the sensitivity and the specificity. However, in some cases the costs of false positives weigh higher than false negatives. In these cases, a $\hat{p}$ with higher sensitivity may be used, while sacrificing specificity.

### 5.3. Receiver Operating Characteristic Curves

Receiver Operating Characteristic (ROC) curves are commonly used in the health science field to evaluate the diagnostic utility and validity of a model. They can be used to represent the model's diagnostic utility across all possible levels of $\hat{p}$ cutoff values. An ROC curve is a plot of a model's or scale's true positives (sensitivity) vs. false positives (1-specificity). An example of curves from two models, one good, one bad can be found in Figure 2.

**Figure 2 Examples of ROC Curves**



With the plot comes a readout of sensitivity and specificity at regular $\hat{p}$ intervals from the logistic regression model. As sensitivity and specificity vary with the $\hat{p}$ values, the ROC can be used to select the $\hat{p}$ value that yields the optimal cutoff point for the purposes of the model. Along with the sensitivity and specificity, overall validity of the model is given at each $\hat{p}$ value. Within the ROC curve, the area under the curve, AUC, is another measure that indicates the accuracy of the test. The AUC is on a scale of 0.5 to 1.0. An AUC of 1.0 signifies a perfect test, an AUC of 0.5 states the test is no more accurate than flipping a coin. The AUC intuitively measures the probability that a neuropathic subject, chosen at random, will be scored higher on the model than a healthy subject, chosen at random. The AUC can also be used to compare the overall diagnostic ability between two tests. In addition, there also exist standard levels of acceptance for AUC values, Swets (1988). Generally, an AUC of .900 or greater represents a model with high accuracy. A more detailed description of AUC can be found in Hanley and McNeil (1982).

There are several instruments for measuring the reliability of a model. These include intra-class correlation, Cronbach's alpha, Yule's Y, kappa, and Pearson's product moment correlation. Descriptions of these can be found in Streiner and Norman (1989) and Bartko (1991). However, not all of these measurements are accurate measures of reliability, (Bartko 1991). For our model, we used kappa, which takes chance agreement into effect. This provides a more accurate validity value than Pearson's product moment correlation coefficient, a instrument used often in reporting validity. Though the values are generally lower than

Pearson's, kappa is a much more valid report of reliability. There are also established standards for kappa, Landis and Koch (1977). Generally, a kappa above .8 is considered near perfect.

## 6. Results

From the questionnaire, questions 22a, 23a, 24a, 25a, 26a, 27a, and 28a were eliminated, as their scoring was not consistent with the other questions. The relationship between the response variable and individual questions was explored using the $X^2$ test and Fisher's exact test. Several variables had $X^2$ tests with the expected counts less than 5. (See appendix 9.2). In these cases Fisher's exact test was used. The results of the analysis can be found in Figure 4.

Figure 3 Analysis of cross-tabulations, with $X^2$ test (X) or Fisher's exact test (F).

| Question | p value | X/F | Question | p value | X/F |
|---|---|---|---|---|---|
| 1 | 0.646 | F | 13 | 1.000 | F |
| 2 | 0.628 | F | 14 | 0.543 | F |
| 3a | 1.000 | F | 15 | 0.127 | F |
| 3b | 0.609 | F | 16 | 0.456 | F |
| 3c | 0.163 | F | 17 | 0.323 | F |
| 3d | 0.295 | F | 18 | 0.743 | F |
| 3e | 0.423 | X | 19 | 1.000 | F |
| 4 | 1.000 | F | 20 | 0.753 | F |
| 5 | 1.000 | F | 21 | 0.585 | F |
| 6 | 0.614 | F | 22 | <.001 | F |
| 7 | 0.638 | F | 23 | 0.498 | F |
| 8 | 0.141 | F | 24 | 0.019 | X |
| 9 | 0.526 | F | 25 | 0.464 | F |
| 10 | 0.235 | X | 26 | 0.410 | F |
| 11 | 1.000 | F | 27 | 0.555 | F |
| 12 | 1.000 | F | 28 | 0.398 | F |

Questions with a p-value of greater than .25 were eliminated as potential model questions. The remaining six questions, Q3C, Q8, Q10, Q15, Q22, and Q24 were analyzed

individually. Question 15 was eliminated, as 95.4% of the subjects responded identically. Question 8 was also problematic, due to zero counts. Further analysis using PROC CATMOD in SAS eliminated Q8 as a potential model question. Question 3C had a problem with zero counts. Therefore, question 3C was collapsed into two levels to compensate for the problem. The new Fisher's exact test p-value was 0.067. Questions 22 and 24 were also examined. Low counts for the highest levels led us to collapse the two highest levels, creating a two level question. The new $X^2$ p-values for 22 and 24 were .001 and .012 respectively.

Question 10 was originally four levels. However, no subjects responded to the two highest levels, thus it was collapsed into a two level question. One subject did not respond to question 10, but the question may be used with the elimination of the subject from the data set. This yielded the chi-squared p-value found in Figure 4, (.235). Numerically, the question was fit to be used as a possible model variable. However, healthy subjects fell more often than neuropathic subjects, which was counter-intuitive to standard reasoning. Thus, I eliminated Q10 as a possible model question.

Thus, three questions were used for the preliminary model. These questions measure difficulty descending stairs (3C), decrease of strength in legs (22), and numbness in legs and feet (24). (Appendix 9.2) The collapsed form of question 3C was used, with two levels: falls in the past year or no falls in the past year. (Appendix 9.4) The collapsed forms of questions 22 and 24 were used as well. These questions had two levels, presence or no presence. (Appendix 9.4).

The three questions were used to create a logistic regression model using SAS PROC LOGISTC. (Appendix 9.6). Since I only had three possible predictor variables, the best subsets

method was used, where all possible subsets of the three predictor variables are fit to determine the optimal model. I did not include one variable models. The results of the models are in Figure 4.

**Figure 4 Logistic Regression Models**

| Model # | Questions Used | Coefficents | Std. Error | Odds Ratio | ROC Analysis | | | p-hat | AUC | H-L GOF |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Sens | Spec | Overall | cutoff | | p-value |
| 1 | Q3C | 1.190 | 1.219 | 3.287 | 92.90 | 37.50 | 72.70 | 0.300 | 0.834 | 0.220 |
| | Q22 | 2.242 | 0.888 | 9.416 | | | | | | |
| | Q24 | 1.284 | 0.907 | 3.610 | | | | | | |
| 2 | Q22 | 2.339 | 0.878 | 10.374 | 92.90 | 37.50 | 72.70 | 0.220 | 0.813 | 0.671 |
| | Q24 | 1.521 | 0.888 | 4.575 | | | | | | |
| 3 | Q3C | 1.579 | 1.192 | 4.850 | 75.00 | 87.50 | 79.50 | 0.400 | 0.799 | 0.111 |
| | Q22 | 2.373 | 0.869 | 10.724 | | | | | | |
| 4 | Q3C | 1.504 | 1.140 | 4.500 | 89.30 | 43.80 | 72.70 | 0.340 | 0.728 | 1.000 |
| | Q24 | 1.540 | 0.815 | 4.666 | | | | | | |

When determining the model, I determined sensitivity weighed higher than specificity, as this is a preliminary screening questionnaire. Since those who are labeled positive on the scale receive further examination, the costs of sending in a healthy diabetic for examination is much lower than that of missing a neuropathic diabetic. Thus I chose as cutoffs for $\hat{p}$ from the ROC analysis that yielded that highest specificity, optimally over 90%, that retained an overall validity of at least 70%. Thus, the model, while giving up some specificity, would allow the scale to identify a very high percentage of those who actually do have neuropathy.

Model 3, using Q3C and Q22, had the highest overall accuracy (79.50%). However, its sensitivity was too low for my purposes (75.00%) and the Hosmer-Lemeshow logistic regression goodness of fit p-value was also low (.111). Model 4, using questions Q3C and Q24, had a perfect Hosmer-Lemeshow logistic regression goodness of fit p-value, (1.000), and good

sensitivity (89.30%) and acceptable overall accuracy (72.70). However, its ROC AUC was lowest of the three models, (.728). Thus, I eliminated models 3 and 4.

This left models 1 and 2, both containing Q22 and Q24, which had the lowest chi-squared p-values of the questions. Model 1 contained Q3C in addition to 22 and 24. The two models were identical in the ROC analysis of sensitivity (92.90%), specificity (37.50%), and overall accuracy (72.70%). Model 1 had a better AUC (.834, .813), but Model 2 had the better Hosmer-Lemeshow logistic regression goodness of fit p-value, (.671, .220). So, neither model was significantly better in prediction utility.

When looking at the cross-tabulations of the two models with neuropathy, I found zero counts in the cross-tabulation of model 1. Model 2 had no zero counts. Thus, as the two models were similar in other aspects, I retained model 2 as our optimal model, which I will now refer to as the model.

**Figure 5 Description of Optimal Logistic Regression Model**

Cross-tabulation for Model 2
(Final Model)

| | | Q24 | | |
|---|---|---|---|---|
| | | 0 | 1 | All |
| Q22 | 0 | 8 | 16 | 24 |
| | neuro | 2 | 8 | 10 |
| | 1 | 2 | 18 | 20 |
| | neuro | 1 | 17 | 18 |
| | All | 10 | 34 | 44 |
| | | 3 | 25 | 28 |

# VALIDATING A HEALTH QUESTIONNAIRE

Classification of Subjects by Model

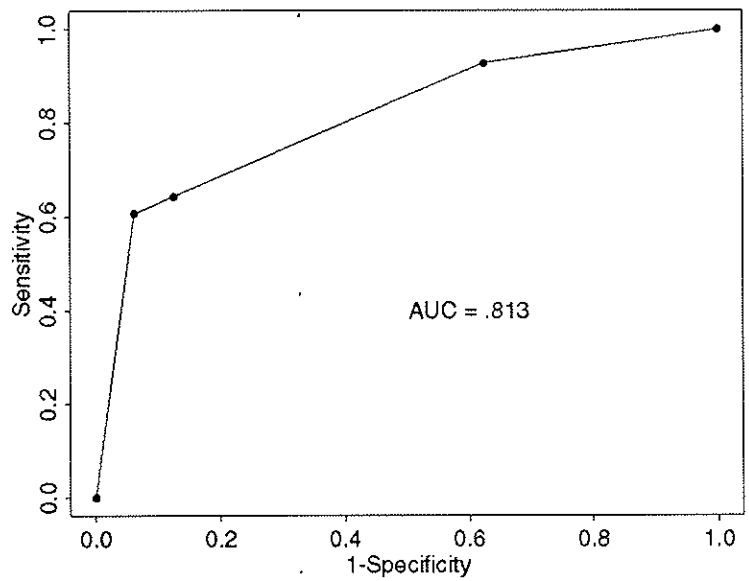
Rows=neuropathics Columns=model classification

| | 0 | 1 | All |
|---|---|---|---|
| 0 | 6 | 10 | 16 |
| 1 | 2 | 26 | 28 |
| All | 8 | 36 | 44 |

Sensitivity= (26/28)=92.90% Specificity= (6/16)=37.50%

Overall accuracy= (32/44)=72.73%

ROC CURVE FOR FINAL MODEL




## Model

$$\ln\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = -1.412 + 2.3393 * strength + 1.5206 * numbers$$

The responses for strength and numbness are coded 0 for no and 1 for yes. So, from the model, a diabetic patient who has a decrease of strength in their feet or legs and numbness in their feet or legs is at about 92% risk for having neuropathy.

Reliability for the model was extremely low (.339). This can be partially attributed to my choice of a high sensitivity. Several of the models had better overall accuracy when higher $\hat{p}$ cutoff values were used. Since the kappa estimate relies heavily on overall accuracy, the 72.73% overall accuracy kept the kappa level low. Also, since I chose such a high value for sensitivity, resulting in a low specificity, the expected counts due to chance were higher than if the two numbers were more balanced. For example, if model 3, with higher overall accuracy and a higher specificity, were used, the kappa improves to an acceptable .586.

I also expect that with increased sample size, the validity and reliability should improve. An increase in sample size should lead to an increase in significant questions, and should greatly reduce the problems with zero counts. The increase in significant questions will allow more models to be fit. With a greater choice of models, one with a better compromise of sensitivity and specificity may found. This will lead to a more valid and reliable model.

## 7. Conclusion

With the data set, I was able to develop a fair model to predict neuropathy within the diabetic subjects. The model is high in sensitivity, 93%, though the specificity is somewhat low, 38.0%. The overall accuracy is acceptable, at 73%. I feel that it can be used clinically as a

preliminary screening device for neuropathy. Future studies in this area may allow the model to be refit, as a larger data set may change the predictive values of the questions, eliminate the problems with zero counts in questions, and allow more questions to be used as predictive variables. I expect the diagnostic capabilities of the questionnaire to increase with the increase in data.

The problem with the zero counts in the cross tabulation analysis of the questions caused me to look closer at the questions themselves. I feel that the number of levels in several of the questions can be reduced. Among all five parts of question 3, no subjects answered the highest level, need assistance, and only two answered the second highest level, very difficult. These questions can be reduced to a two level question, with adjustment of the wording. Question 10 can also be changed, as again no subject answered the two highest levels of the four level question. The same is true for questions 22 and 24. All questions were significant when the number of levels was reduced. A complete summary of the questions with zero counts can be found in Appendix 9.4.

While the levels of several questions may be reduced, cross-validation of the results is needed before any major reconstruction should be made to the questionnaire. I look forward to monitoring the clinical utility of the questionnaire.

## 8. References

Agresti, A. (1990) Categorical Data Analysis. New York: John Wiley and Sons.

Agresti, A. (1996) An Introduction to Categorical Data Analysis. New York: John Wiley and Sons

Bartko, J. J. (1991). Measurement and reliability: statistical thinking considerations. Schizophrenia Bulletin, 17(3), 483.

Bohn, M.J., Babor, T.F. & Kranzler, H.R. (1995). The alcohol use disorders identification test (AUDIT): validation of a screening instrument for use in medical settings. Journal of Studies on Alcohol, July, 423.

Cavanagh, P.R., Derr, J.A., Ulbrecht, J.S:, Maser, R.E. & Orchard, T.J. (1992). Problems with gait and posture in neuropathic patients with insulin-dependent diabetes mellitus. Diabetic Medicine, 9, 469-474.

Chen, W.J., Faraone, S.V., Biederman, & Tsuang, T.S. (1994). diagnostic accuracy of the child behavior checklist scales for attention-deficit hyperactivity disorder: a receiver-operating characteristic analysis. Journal for Consulting and Clinical Psychology, 62(5), 1017.

Hanley. J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristc curve. Diagnostic Radiology; 143(1):29-36.

Hosmer, D.W. & Lemeshow, S. (1989). Applied Logistic Regression. New York: John Wiley and Sons.

Kleinbaum, D.G. (1992). Logistic Regression: A Self-Learning Text. New York: Springer-

Verlag.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics; 33:159-174.

McDowell, I. & Newell, C. (1996). Measuring Health 2nd Edition. New York: Oxford University Press.

Orchard, T.J., Dorman, J.S., Maser R.E., Blecker, D.J., Drash, A.L., Ellis, D. et al. (1990). Prevelence of complications in IDDM by sex and duration. Diabetes; 39:1116-1124

Pirat J. (1979). Diabetes mellitus and its degenerative complications: A prospective study of 4,400 patients observed between 1947 and 1973. Diabetes Care; 2:168-188 and 252-263.

Streiner, D.L. & Norman, G.R. (1989). Health Measurement Scales. New York: Oxford University Press.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. Science; 240:1285-1293

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental tool in clinical medicine. Clinical Chemistry, 39(4), 561-577.

# 9. Appendix

## 9.1. Synopsis

Since my work was not done at St. John's, I would like to give a synopsis of what I did at Penn State.

I had no knowledge of neuropathy, logistic regression, ROC curves, or most of what this thesis involved. When I got to Penn State, I was tentatively assigned to a different project involving writing a long SAS program. This was not what I wanted to do for two months, so I looked for something different. I attended an informational session on the diabetes project, and it sounded interesting. So, I switched projects.

After joining the project, I met with the people at CELOS to get up to speed on the project. I was given a tour of the facilities, and got an understanding of the type of work done at CELOS. The project I was involved with was an offshoot of a kinesiology graduate student's dissertation work. The people at CELOS and the Statistical Consulting Center had not gotten around to working on the project, so the opportunity was open for me to work on it. The people at the Consulting Center did not know how to approach the project, so they let me go at it and try to discover a way to model the presence of neuropathy.

As mentioned before, I had no previous knowledge of neuropathy or questionnaire validation, and I had a limited knowledge of statistics (one year of undergraduate courses). So, I spent a lot of time in the library looking up research articles, papers, and journals to what kind of

previous work had been done in the area. The Patee Library at PSU has an extensive collection of research journals, so I was able to find a large number of sources. As I was doing my library research, I was experimenting with ways to label the subjects, most of them pretty crude. I tried a lot of different things, and most of them didn't work. I did the basic exploratory data analysis, as I knew what I was doing. After sifting through the papers, and consulting with Dr. Janice Derr and John O'Gorman in the Consulting Center, I decided that logistic regression would be the best tool to use to create my model.

So, I did a lot of reading and learned the basics of logistic regression. I also found ROC analysis in several papers. So, I researched ROC analysis and discovered how to use it. As I was doing this, I learned how to use SAS and Minitab to do logistic regression. John O'Gorman helped me write the SAS programs needed to do the analysis. From the SAS programs, I was able to generate the potential models and determine the estimates of the coefficients. I did more library research to determine how to compare the potential models. With about a week left, I came up with my final model.

So, in the ten weeks at Penn State, I started with no knowledge of the project and came to a good understanding of modeling from a questionnaire. I spent a lot of time doing library research, reading up on methods, and doing computer analysis. Eventually, everything came together, and made sense. THE END

## *9.2. Questionnaire*

### POSTURE PROJECT QUESTIONNAIRE (MASTER COPY)

*This is a self-administered questionnaire. Answer each question with what you feel is the most appropriate answer. Note that since this test is self-administered, the testers were instructed not to give you any additional information.*

### Part I: Fall and activity level survey

**Section A.** *The following three questions are related to your current activity level.*

***For each question choose the best answer.***

1. Which one of the following describes your current recreational category best?
   a. no particular exercise ❑ [1]
   b. moderately active (walking/swimming/cycling/gardening) ❑ [2]
   c. very athletic (running/racquet events/field sports) ❑ [3]

2. Which one of the following describes your current occupational/work category best?
   a. not currently employed (no housework) ❑ [1]
   b. not currently employed (do housework, yard work, etc.) ❑ [2]
   c. lightly strenuous work (desk work, secretary, receptionist) ❑ [3]
   d. moderately strenuous work (clerk, salesperson, etc.) ❑ [4]
   e. very strenuous work (construction work, mailman, etc.) ❑ [5]

3. How would you rate your ability to perform the following tasks?

***Select one answer for each of the activities listed.***

a. Standing still:

| ❑ no difficulty [0] | ❑ somewhat difficult [1] | ❑ very difficult [2] | ❑ need assistance [3] |
|---|---|---|---|

b. Walking:

| ❑ no difficulty [0] | ❑ somewhat difficult [1] | ❑ very difficult [2] | ❑ need assistance [3] |
|---|---|---|---|

c. Stair descent:

| ❑ no difficulty [0] | ❑ somewhat difficult [1] | ❑ very difficult [2] | ❑ need assistance [3] |
|---|---|---|---|

d. Stair ascent:

| ❑ no difficulty [0] | ❑ somewhat difficult [1] | ❑ very difficult [2] | ❑ need assistance [3] |
|---|---|---|---|

e. Getting up out of a low chair:

| ❑ no difficulty [0] | ❑ somewhat difficult [1] | ❑ very difficult [2] | ❑ need assistance [3] |
|---|---|---|---|

**Section B.** *The following questions are designed to provide us with information about problems or difficulties that you may have encountered while standing or walking.*

Slipping or tripping

4. Do you worry about slipping or tripping when you walk?

❑ Yes [1] ❑ No [0]

5. Over the last four years, have you changed the way you stand or walk to prevent slips or trips?

❑ Yes [1] ❑ No [0]

6. Over the last four years, have you changed your activity patterns in any other way to prevent slips or trips ?

❑ Yes [1] ❑ No [0]

Unsteadiness/Balance

7. Do you worry about unsteadiness or loosing your balance when you are standing or walking?

❑ Yes [1] ❑ No [0]

8. Over the last four years, have you changed the way you stand or walk to prevent feeling unsteady?

❑ Yes [1] ❑ No [0]

9. Over the last four years, have you changed your activity patterns in any other way to prevent feeling unsteady?

❑ Yes [1] ❑ No [0]

Falls

10. How many times have you fallen in the last year?

❑ not at all [0] ❑ 1 to 5 times [1] ❑ 5 to 10 times [2] ❑ more than [3] 10 times

11. Do you worry about falling?

❑ Yes [1] ❑ No [0]

12. Over the past four years, have you changed the way you stand or walk to prevent falling?

❑ Yes [1] ❑ No [0]

13. Over the last four years, have you changed your activity patterns in any other way to prevent falling?

❑ Yes [1] ❑ No [0]

Injuries

14. Have you fractured a bone during the last four years as the result of a slip, trip or fall?

❑ Yes [1] ❑ No [0]

15. Have you sprained your ankle in the last four years as the result of a slip, trip or fall?

❑ Yes [1] ❑ No [0]

16. Have you sustained any cuts or bruises in the last four years as the result of a slip, trip or fall?

❑ Yes [1] ❑ No [0]

**Section C.** *The following questions are designed to provide us with information about problems that you may have encountered while standing or walking in some special circumstances.*

17. How comfortable do you feel going down a flight of stairs with which you are unfamiliar?

❑ very safe [0] ❑ safe [1] ❑ a little unsafe [2] ❑ very unsafe [3]

18. How comfortable do you feel standing in the shower without the use of supports?

❑ very safe [0] ❑ safe [1] ❑ a little unsafe [2] ❑ very unsafe [3]

19. How comfortable do you feel standing or walking on uneven ground?

❑ very safe [0] ❑ safe [1] ❑ a little unsafe [2] ❑ very unsafe [3]

20. How comfortable do you feel standing or walking in a dimly lit room with which you are unfamiliar?

❑ very safe [0] ❑ safe [1] ❑ a little unsafe [2] ❑ very unsafe [3]

21. How comfortable do you feel walking across an open space such as a field or parking lot?

❑ very safe [0] ❑ safe [1] ❑ a little unsafe [2] ❑ very unsafe [3]

## Part II: Perceived sensory and motor function

*The following questions are designed to provide us with information about how you feel about the level of function in your legs and feet.*

*For the following question, please answer yes, sometimes or no/never. If you answer yes or sometimes, please qualify your answer by indicating the level of severity of your complaint by putting a vertical line on a scale of 1 to 10. A rating of 1 corresponds to a* ***minor*** *problem which means that it does not interfere with daily activities. A rating of 10 corresponds to a* ***major*** *problem which means that it does significantly interfere with daily activities, for example; housework, gardening, walking, work, etc.*

22. Do you currently have a decrease in the strength of your legs or feet that is out of proportion with any general changes in your overall strength?

❑ yes [2] ❑ sometimes [1] ❑ no [0]

***If yes or sometimes, indicate the severity of this loss of strength on a scale from 1 to 10:***

mild_____severe

1 10

23. Do your toes or the soles of your shoes tend to drag or catch on the floor when you are walking?

❑ yes [2] ❑ sometimes [1] ❑ never [0]

***If yes or sometimes, indicate the severity of this problem on a scale from 1 to 10:***

mild_____severe

1 10

24. Do you have numbness in your feet?

❑ yes [2] ❑ sometimes [1] ❑ never [0]

***If yes or sometimes, indicate the severity of this problem on a scale from 1 to 10:***

mild_____severe

1 10

25. Do you feel tingling, "pins and needles", burning, deep itching or other unusual sensation in your feet?

❑ yes [2] ❑ sometimes [1] ❑ never [0]

***If yes or sometimes, indicate the severity of this tingling, "pins and needles", burning, deep itching or other unusual sensation on a scale from 1 to 10:***

mild_____severe

1 10

26. Do you have increased sensitivity of your feet, such that shoes or objects touching your feet lightly (for example: bed sheets at night) might bother you by causing pain or unpleasant sensation?

❑ yes [2] ❑ sometimes [1] ❑ never [0]

***If yes or sometimes, indicate the severity of this problem on a scale from 1 to 10:***

mild_____severe

1 10

27. Do you have pain (aching, dull, lancing, shooting) in your legs or feet?

❑ yes [2] ❑ sometimes [1] ❑ never [0]

***If yes or sometimes, indicate the severity of this problem on a scale from 1 to 10:***

mild_____severe

1 10

28. Do you have problems with coordination of your legs and/or feet (i.e., a clumsiness while walking, stumbling)?

❑ yes [2] ❑ sometimes [1] ❑ never [0]

***If yes or sometimes, indicate the severity of this problem on a scale from 1 to 10:***

mild_____severe

1 10

## 9.3. Chi-Square Tests and Contingency Tables

STATISTICS FOR TABLE OF Q1 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 1.280 | 0.527 |
| Fisher's Exact Test (2-Tail) | | | 0.646 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q2 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 4 | 2.937 | 0.568 |
| Fisher's Exact Test (2-Tail) | | | 0.628 |

Sample Size = 44
WARNING: 60% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q3A BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.013 | 0.910 |
| Fisher's Exact Test (Left) | | | 0.753 |
| (Right) | | | 0.704 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q3B BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 1.856 | 0.395 |
| Fisher's Exact Test (2-Tail) | | | 0.609 |

Sample Size = 44
WARNING: 67% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q3C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 3.935 | 0.140 |
| Fisher's Exact Test (2-Tail) | | | 0.163 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less

than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q3CCOLL BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 3.887 | 0.049 |
| Fisher's Exact Test (Left) | | | 0.995 |
| (Right) | | | 0.050 |
| (2-Tail) | | | 0.067 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q3D BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 3.054 | 0.217 |
| Fisher's Exact Test (2-Tail) | | | 0.295 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q3E BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.642 | 0.423 |
| Fisher's Exact Test (Left) | | | 0.868 |
| (Right) | | | 0.315 |
| (2-Tail) | | | 0.534 |

Sample Size = 44

STATISTICS FOR TABLE OF Q4 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.035 | 0.851 |
| Fisher's Exact Test (Left) | | | 0.557 |
| (Right) | | | 0.705 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q5 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.074 | 0.786 |
| Fisher's Exact Test (Left) | | | 0.533 |
| (Right) | | | 0.744 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q6 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.354 | 0.552 |
| Fisher's Exact Test (Left) | | | 0.463 |
| (Right) | | | 0.871 |
| (2-Tail) | | | 0.614 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q7 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.653 | 0.419 |
| Fisher's Exact Test (Left) | | | 0.910 |
| (Right) | | | 0.392 |
| (2-Tail) | | | 0.638 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q8 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 3.223 | 0.073 |
| Fisher's Exact Test (Left) | | | 1.000 |
| (Right) | | | 0.090 |
| (2-Tail) | | | 0.141 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q9 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.197 | 0.274 |
| Fisher's Exact Test (Left) | | | 1.000 |
| (Right) | | | 0.400 |
| (2-Tail) | | | 0.526 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q11 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.065 | 0.798 |
| Fisher's Exact Test (Left) | | | 0.724 |

| | | |
|---|---|---|
| | (Right) | 0.544 |
| | (2-Tail) | 1.000 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q12 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.218 | 0.640 |
| Fisher's Exact Test (Left) | | | 0.812 |
| (Right) | | | 0.496 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

The SAS System

STATISTICS FOR TABLE OF Q13 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.246 | 0.620 |
| Fisher's Exact Test (Left) | | | 0.849 |
| (Right) | | | 0.537 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q14 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.278 | 0.258 |
| Fisher's Exact Test (Left) | | | 0.296 |
| (Right) | | | 0.958 |
| (2-Tail) | | | 0.543 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q15 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 3.667 | 0.056 |
| Fisher's Exact Test (Left) | | | 0.127 |
| (Right) | | | 1.000 |
| (2-Tail) | | | 0.127 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q16 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.040 | 0.308 |

| | | |
|---|---|---|
| Fisher's Exact Test | (Left) | 0.256 |
| | (Right) | 0.917 |
| | (2-Tail) | 0.456 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q17 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 2.424 | 0.298 |
| Fisher's Exact Test (2-Tail) | | | 0.323 |

Sample Size = 44
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q19 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 3 | 0.599 | 0.897 |
| Fisher's Exact Test (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q20 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 3 | 1.748 | 0.626 |
| Fisher's Exact Test (2-Tail) | | | 0.753 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q21 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 1.833 | 0.400 |
| Fisher's Exact Test (2-Tail) | | | 0.585 |

Sample Size = 44
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q22 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 11.099 | 0.004 |
| Fisher's Exact Test (2-Tail) | | | 3.31E-03 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q23 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 1.834 | 0.400 |
| Fisher's Exact Test (2-Tail) | | | 0.498 |

Sample Size = 44
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q24 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 7.916 | 0.019 |
| Fisher's Exact Test (2-Tail) | | | 0.023 |

Sample Size = 44

STATISTICS FOR TABLE OF Q25 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 2.161 | 0.339 |
| Fisher's Exact Test (2-Tail) | | | 0.464 |

Sample Size = 44
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q26 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 2.110 | 0.348 |
| Fisher's Exact Test (2-Tail) | | | 0.410 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q27 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 1.389 | 0.499 |
| Fisher's Exact Test (2-Tail) | | | 0.555 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

STATISTICS FOR TABLE OF Q28 BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 2.154 | 0.341 |
| Fisher's Exact Test (2-Tail) | | | 0.398 |

Sample Size = 44
WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

## 9.4. Edited Questions

22. Do you currently have a decrease in the strength of your legs or feet that is out of proportion with any general changes in your overall strength?

❑ yes [1] ❑ no [0]

24. Do you ever have numbness in your feet?

❑ yes [1] ❑ never [0]

## 9.5. Collapsed Question Analysis

The SAS System

TABLE OF Q3AC BY NEUROP

| Q3AC | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no difficulty | 15<br>93.75 | 26<br>92.86 | 41 |
| >= some difficulty | 1<br>6.25 | 2<br>7.14 | 3 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q3AC BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.013 | 0.910 |
| Fisher's Exact Test (Left) | | | 0.753 |
| (Right) | | | 0.704 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q3BC BY NEUROP

| Q3BC | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no difficulty | 15<br>93.75 | 22<br>78.57 | 37 |
| >= some difficulty | 1<br>6.25 | 6<br>21.43 | 7 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q3BC BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.753 | 0.185 |
| Fisher's Exact Test (Left) | | | 0.969 |
| (Right) | | | 0.188 |
| (2-Tail) | | | 0.393 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q3CC BY NEUROP

| Q3CC | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no difficulty | 15<br>93.75 | 19<br>67.86 | 34 |
| >= some difficulty | 1<br>6.25 | 9<br>32.14 | 10 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q3CC BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 3.887 | 0.049 |
| Fisher's Exact Test (Left) | | | 0.995 |
| (Right) | | | 0.050 |
| (2-Tail) | | | 0.067 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q3DC BY NEUROP

| Q3DC | NEUROP(Neuropathy) |
|---|---|
| Frequency | |

| Col Pct | Non Neur | Neur | Total |
|---|---|---|---|
| no difficulty | 14<br>87.50 | 18<br>64.29 | 32 |
| >= some difficulty | 2<br>12.50 | 10<br>35.71 | 12 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q3DC BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 2.766 | 0.096 |
| Fisher's Exact Test | (Left) | | | 0.982 |
| | (Right) | | | 0.092 |
| | (2-Tail) | | | 0.160 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q3EC BY NEUROP

Q3EC NEUROP(Neuropathy)

Frequency

| Col Pct | Non Neur | Neur | Total |
|---|---|---|---|
| no difficulty | 10<br>62.50 | 14<br>50.00 | 24 |
| >= some difficulty | 6<br>37.50 | 14<br>50.00 | 20 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q3EC BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 0.642 | 0.423 |
| Fisher's Exact Test | (Left) | | | 0.868 |
| | (Right) | | | 0.315 |
| | (2-Tail) | | | 0.534 |

Sample Size = 44

The SAS System

TABLE OF Q10C BY NEUROP

Q10C NEUROP(Neuropathy)

Frequency,

| Col Pct | Non Neur | Neur | Total |
|---|---|---|---|
| 0 falls | 9<br>56.25 | 20<br>71.43 | 29 |
| >0 falls | 7<br>43.75 | 8<br>28.57 | 15 |

```
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total          16       28       44
```

STATISTICS FOR TABLE OF Q10C BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 1.044 | 0.307 |
| Fisher's Exact Test | (Left) | | | 0.244 |
| | (Right) | | | 0.911 |
| | (2-Tail) | | | 0.340 |

Sample Size = 44

The SAS System

TABLE OF Q17C BY NEUROP

| Q17C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| very safe or safe | 13<br>81.25 | 18<br>64.29 | 31 |
| a little unsafe or unsafe | 3<br>18.75 | 10<br>35.71 | 13 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q17C BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 1.408 | 0.235 |
| Fisher's Exact Test | (Left) | | | 0.940 |
| | (Right) | | | 0.201 |
| | (2-Tail) | | | 0.314 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q18C BY NEUROP

| Q18C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| very safe or safe | 15<br>93.75 | 24<br>85.71 | 39 |
| a little unsafe or unsafe | 1<br>6.25 | 4<br>14.29 | 5 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q18C BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 0.653 | 0.419 |

| | | |
|---|---|---|
| Fisher's Exact Test | (Left) | 0.910 |
| | (Right) | 0.392 |
| | (2-Tail) | 0.638 |

Sample Size = 44
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q19C BY NEUROP

| Q19C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| very safe or safe | 12<br>75.00 | 20<br>71.43 | 32 |
| a little unsafe or unsafe | 4<br>25.00 | 8<br>28.57 | 12 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q19C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 0.065 | 0.798 |
| Fisher's Exact Test (Left) | | | 0.724 |
| (Right) | | | 0.544 |
| (2-Tail) | | | 1.000 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q20C BY NEUROP

| Q20C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| very safe or safe | 13<br>81.25 | 18<br>64.29 | 31 |
| a little unsafe or unsafe | 3<br>18.75 | 10<br>35.71 | 13 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q20C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.408 | 0.235 |
| Fisher's Exact Test (Left) | | | 0.940 |
| (Right) | | | 0.201 |
| (2-Tail) | | | 0.314 |

Sample Size = 44

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q21C BY NEUROP

| Q21C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| very safe or safe | 15<br>93.75 | 28<br>100.00 | 43 |
| a little unsafe or unsafe | 1<br>6.25 | 0<br>0.00 | 1 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q21C BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 1.791 | 0.181 |
| Fisher's Exact Test | (Left) | | | 0.364 |
| | (Right) | | | 1.000 |
| | (2-Tail) | | | 0.364 |

Sample Size = 44

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q22C BY NEUROP

| Q22C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no | 14<br>87.50 | 10<br>35.71 | 24 |
| at least sometimes | 2<br>12.50 | 18<br>64.29 | 20 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q22C BY NEUROP

| Statistic | | DF | Value | Prob |
|---|---|---|---|---|
| Chi-Square | | 1 | 11.013 | 0.001 |
| Fisher's Exact Test | (Left) | | | 1.000 |
| | (Right) | | | 9.59E-04 |
| | (2-Tail) | | | 1.36E-03 |

Sample Size = 44

The SAS System

TABLE OF Q23C BY NEUROP

| Q23C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no | 10<br>62.50 | 13<br>46.43 | 23 |
| at least sometimes | 6<br>37.50 | 15<br>53.57 | 21 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q23C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.054 | 0.305 |
| Fisher's Exact Test (Left) | | | 0.910 |
| (Right) | | | 0.239 |
| (2-Tail) | | | 0.360 |

Sample Size = 44

The SAS System

TABLE OF Q24C BY NEUROP

| Q24C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no | 7<br>43.75 | 3<br>10.71 | 10 |
| at least sometimes | 9<br>56.25 | 25<br>89.29 | 34 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q24C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 6.327 | 0.012 |
| Fisher's Exact Test (Left) | | | 0.998 |
| (Right) | | | 0.017 |
| (2-Tail) | | | 0.022 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q25C BY NEUROP

| Q25C | NEUROP(Neuropathy) | | |
|---|---|---|---|
| Frequency<br>Col Pct | Non Neur | Neur | Total |
| no | 6<br>37.50 | 5<br>17.86 | 11 |

| | | | |
|---|---|---|---|
| at least sometim es | 10<br>62.50 | 23<br>82.14 | 33 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q25C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 2.095 | 0.148 |
| Fisher's Exact Test (Left) | | | 0.963 |
| (Right) | | | 0.139 |
| (2-Tail) | | | 0.169 |

Sample Size = 44
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

The SAS System

TABLE OF Q26C BY NEUROP

| Q26C<br>Frequency<br>Col Pct | NEUROP(Neuropathy)<br>Non Neur | Neur | Total |
|---|---|---|---|
| no | 12<br>75.00 | 16<br>57.14 | 28 |
| at least sometim es | 4<br>25.00 | 12<br>42.86 | 16 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q26C BY NEUROP

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 1.403 | 0.236 |
| Fisher's Exact Test (Left) | | | 0.937 |
| (Right) | | | 0.196 |
| (2-Tail) | | | 0.333 |

Sample Size = 44

The SAS System

TABLE OF Q27C BY NEUROP

| Q27C<br>Frequency<br>Col Pct | NEUROP(Neuropathy)<br>Non Neur | Neur | Total |
|---|---|---|---|
| no | 6<br>37.50 | 7<br>25.00 | 13 |
| at least sometim es | 10<br>62.50 | 21<br>75.00 | 31 |
| Total | 16 | 28 | 44 |

STATISTICS FOR TABLE OF Q27C BY NEUROP

```
Statistic                       DF      Value        Prob
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Chi-Square                       1     0.764       0.382

Fisher's Exact Test (Left)                         0.888
                    (Right)                        0.295
                    (2-Tail)                       0.496

Sample Size = 44
WARNING: 25% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.
```

```
                    The SAS System

                TABLE OF Q28C BY NEUROP

     Q28C               NEUROP(Neuropathy)

     Frequency          ,
     Col Pct            ,Non Neur,Neur     ,  Total
     ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
     no                 ,     10 ,     16 ,     26
                        ,  62.50 ,  57.14 ,
     ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
     at least sometim   ,      6 ,     12 ,     18
     es                 ,  37.50 ,  42.86 ,
     ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
     Total                    16       28        44


        STATISTICS FOR TABLE OF Q28C BY NEUROP

Statistic                       DF      Value        Prob
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Chi-Square                       1     0.121       0.728

Fisher's Exact Test (Left)                         0.746
                    (Right)                        0.491
                    (2-Tail)                       0.761

Sample Size = 44
```

# 9.6. SAS Programs

## 9.6.1. Logistic Regression

```
*******************************************************************************;
* Date:                 7-27-97                                                 ;
* Last modified:        8-06-97                                                 :
* Programmer:           John O'Gorman and Matt Maurer                              ;
* Project:              Posture-Cross Validation                                ;
* Purpose:              Logistic Regression                                     ;
* Program name:         Logreg.sas                                              ;
* Program location:     F:\Research\CELOS-Posture Cross Validation\Examples\    ;
* Data location:        F:\Research\CELOS-Posture Cross Validation\Examples\AnswersOnly.csv;
* Output file:          LogisticReg.lst                                         ;
* Reference:            Chapter 16 of SAS/STAT Software Changes and Enhancements ;
*                       through Release 6.11                                    ;
*******************************************************************************;
```

# VALIDATING A HEALTH QUESTIONNAIRE

```
options nodate nonumber linesize=64 pagesize=60;

* To read an Excel file into SAS, first delete out the column names and save it as  ;
* a CSV (comma delimited) file.  The line below tells SAS the location of the data file;
* and give it an internal name called in                                              ;

filename in 'F:\Scc Research\CELOS-Posture Cross Validation\Data\AnswersOnly.csv';

*************************************************************************************;
* Formats for categorical variable;
*************************************************************************************;

proc format;
    value neurfmt  0 = '0-Non Neuropathic'
                   1 = '1-Neuropathic';


*************************************************************************************;
* Read in the AnswersOnly.csv data file                                              ;
*************************************************************************************;

data answers;
*   The name in refers to the name given in the filename statement above;
*   The delimiter=',' tells SAS that observations on each line are separated by a comma;
    infile in delimiter=',';
    input subject Q1 Q2 Q3a Q3b Q3c Q3d Q3e Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18
Q19 Q20 Q21 Q22 Q22a Q23 Q23a Q24 Q24a Q25 Q25a Q26 Q26a Q27 Q27a Q28 Q28a group;
*   Drop out the non diabetics;
    if group = 1 then delete;
*   Define the neuropathy variable;
    if group > 2 then neurop = 1;
        else neurop = 0;
*   Delete subjects 11 and 48 who did not fill out the questionnaire;
    if subject = 11 or subject = 48 then delete;
*   Assigns the neurop variable the format described in proc format;
*   attrib neurop label = 'Neuropathy' format = neurfmt.;

*   Drop out questions with missing values;
data complete;
    set answers (drop = Q18);

*   Create data reduction categories;
data collapse;
    set complete;
         if Q3c > 0 then q3ccol1 = 1;
           else q3ccol1 = 0;
         if q22 > 0 then q22b = 1;
           else q22b = 0;
         if q24 > 0 then q24b = 1;
           else q24b = 0;
run;


*************************************************************************************;
* Run the logistic regression                                                        ;
*************************************************************************************;

* You use the DESCENDING option when the response variable is coded 1 as an event and;
* 0 as a nonevent.  SAS will then treat 1 as the first ordered value and 0 as the ;
* second ordered value.  This is necessary since SAS will model the probability of;
* the first ordered value;
proc logistic data=collapse descending;
* The lackfit option performs the Hosmer-Lemeshow goodness of fit test;
* The ctable option produced information on specificity, sensitivity, false-positive;
*   rates and false-negative rates;
* The outroc=rocdata will store information about the Receiver Operated Curve;
*   in a data set called rocdata;
    model neurop =  Q22b Q24b  / lackfit ctable outroc=rocdata;
* Store the predicted values in a dataset called pred;
    output out=pred p=phat;
    title 'Logistic Regression';

*************************************************************************************;
* Print out the predicted probabilities in the pred data set;
*************************************************************************************;
```

```
proc print data=pred;
    title 'Predicted Probabilities';

*******************************************************************************;
* Determine the contents of the rocdata data set;
*******************************************************************************;

proc contents data=rocdata;
    title 'Contents of rocdata data set';

*******************************************************************************;
* Plot sensitivity (_SENSIT_) versus 1-specificity (_1MSPEC_);
*******************************************************************************;

proc plot data=rocdata;
    plot _SENSIT_*_1MSPEC_;
    title 'Traditional Receiver Operated Curve';
run;
```

### 9.6.2. Data Analysis

```
*******************************************************************************;
* Date:                7-16-97                                                 ;
* Last modified:       7-16-97                                                 ;
* Programmer:          John O'Gorman                                           ;
* Project:             Posture-Cross Validation                                ;
* Purpose:              PROC CATMOD analysis;
* Program name:        Sparsedata.sas                                         ;
* Program location:    F:\Research\CELOS-Posture Cross Validation\Examples\    ;
* Data location:       F:\Research\CELOS-Posture Cross Validation\Examples\LogisticReg.csv;
* Output file:         LogisticReg.lst                                         ;
* Reference:           Chapter 16 of SAS/STAT Software Changes and Enhancements ;
*                      through Release 6.11                                    ;
*******************************************************************************;

options nodate nonumber linesize=64 pagesize=60;

* To read an Excel file into SAS, first delete out the column names and save it as  ;
* a CSV (comma delimited) file.  The line below tells SAS the location of the data file;
* and give it an internal name called in                                          ;

filename in 'F:\Scc Research\CELOS-Posture Cross Validation\Data\AnswersOnly.csv';

*******************************************************************************;
* Formats for categorical variable;
*******************************************************************************;

proc format;
    value neurfmt  0 = '0-Non Neuropathic'
                   1 = '1-Neuropathic';


*******************************************************************************;
* Read in the AnswersOnly.csv data file                                        ;
*******************************************************************************;

data answers;
*   The name in refers to the name given in the filename statement above;
*   The delimiter=',' tells SAS that observations on each line are separated by a comma;
    infile in delimiter=',';
    input subject Q1 Q2 Q3a Q3b Q3c Q3d Q3e Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18
Q19 Q20 Q21 Q22 Q22a Q23 Q23a Q24 Q24a Q25 Q25a Q26 Q26a Q27 Q27a Q28 Q28a group;
*   Define the neuropathy variable;
    if group > 2 then neurop = 1;
        else neurop = 0;
*   Delete subjects 11 and 48 who did not fill out the questionnaire;
    if subject = 11 or subject = 48 then delete;
*   Assigns the neurop variable the format described in proc format;
*   attrib neurop label = 'Neuropathy' format = neurfmt.;

proc print data = answers;

data sparse;
```

```
    input q8 neurop count;
    cards;
0 0 16
0 1 23
1 0 0
1 1 5
;

data add1;
    set sparse;
    count = count + .5;

proc catmod data=add1;
    weight count;
    model neurop = q8;
    title 'Add 0.5 to each cell';

data add2;
    set sparse;
    count = count + .05;

proc catmod data=add2;
    weight count;
    model neurop = q8;
    title 'Add 0.05 to each cell';

data add3;
    set sparse;
    count = count + .005;

proc catmod data=add3;
    weight count;
    model neurop = q8;
    title 'Add 0.005 to each cell';

data add4;
    set sparse;
    count = count + .0005;

proc catmod data=add4;
    weight count;
    model neurop = q8;
    title 'Add 0.0005 to each cell';

data add5;
    set sparse;
    count = count + .00005;

proc catmod data=add5;
    weight count;
    model neurop = q8;
    title 'Add 0.00005 to each cell';

run;
```

## 9.6.3. Cross Tabulations and Collapse

```
*******************************************************************************;
* Date:                7-22-97                                                 ;
* Last modified:       8-2-97 (by John O'Gorman)                               ;
* Programmers:         Matt Maurer and John O'Gorman                           ;
* Project:             Posture-Cross Validation                                ;
* Purpose:             Cross tabs for questions which collapsed categories     ;
* Directory:           \\Keeper\Scc\Scc Research\CELOS-Posture Cross Validation ;
* Program:             Programs\Cross Tabs - Collapse.sas                      ;
* Data:                Data\AnswersOnly.csv                                    ;
* Output file:         Output\Cross Tabs - Collapse.lst                        ;
*******************************************************************************;

options nodate nonumber linesize=64 pagesize=60;

* To read an Excel file into SAS, first delete out the column names and save it as  ;
```

```
* a CSV (comma delimited) file.  The line below tells SAS the location of the data file;
* and give it an internal name called in                                              ;

filename in '\\Keeper\Scc\Scc Research\CELOS-Posture Cross Validation\Data\AnswersOnly.csv';

*************************************************************************************;
* Formats for categorical variables;
*************************************************************************************;

proc format;
    value neurfmt  0 = 'Non Neur'
                   1 = 'Neur';
    value yesnofmt 0 = 'No'
                   1 = 'Yes';
    value Q3Cfmt   0 = 'no difficulty'
                   1 = '>= some difficulty';
    value Q10Cfmt  0 = '0 falls'
                   1 = '>0 falls';
    value SectCfmt 0 = 'very safe or safe'
                   1 = 'a little unsafe or unsafe';
    value PartIIfmt 0 = 'no'
                    1 = 'at least sometimes';


*************************************************************************************;
* Read in the AnswersOnly.csv data file                                              ;
*************************************************************************************;

data answers;
*   The name in refers to the name given in the filename statement above;
*   The delimiter=',' tells SAS that observations on each line are separated by a comma;
    infile in delimiter=',';
    input subject Q1 Q2 Q3a Q3b Q3c Q3d Q3e Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18
Q19 Q20 Q21
          Q22 Q22a Q23 Q23a Q24 Q24a Q25 Q25a Q26 Q26a Q27 Q27a Q28 Q28a group;
*   Drop out the non diabetics;
    if group = 1 then delete;
*   Define the neuropathy variable;
    if group > 2 then neurop = 1;
       else neurop = 0;
*   Delete subjects 11 and 48 who did not fill out the questionnaire;
    if subject = 11 or subject = 48 then delete;
*   Assigns the neurop variable the format described in proc format;
    attrib neurop label = 'Neuropathy' format = neurfmt.;

*************************************************************************************;
* Collapse categories for some of the questions                                      ;
*************************************************************************************;

data collapse;
    set answers;
*   For Q3aC-Q3eC, 0 is no difficulty and 1 is at least some difficulty;
    if Q3a > 0 then Q3aC = 1;
        else Q3aC = 0;
    if Q3b > 0 then Q3bC = 1;
        else Q3bC = 0;
    if Q3c > 0 then Q3cC = 1;
        else Q3cC = 0;
    if Q3d > 0 then Q3dC = 1;
        else Q3dC = 0;
    if Q3e > 0 then Q3eC = 1;
        else Q3eC = 0;
*   For Q10C, 0 is 0 falls, and 1 is at least 1 fall;
    if Q10 > 0 then Q10C = 1;
        else Q10C = 0;
*   For Q17C-Q21C, 0 is very safe or safe and 1 is a little unsafe or very unsafe;
    if Q17 > 1 then Q17C = 1;
        else Q17C = 0;
    if Q18 > 1 then Q18C = 1;
        else Q18C = 0;
    if Q19 > 1 then Q19C = 1;
        else Q19C = 0;
    if Q20 > 1 then Q20C = 1;
        else Q20C = 0;
    if Q21 > 1 then Q21C = 1;
        else Q21C = 0;
```

```
*   For Q22C-Q28C, 0 is no and 1 is at least sometimes;
    if Q22 > 0 then Q22C = 1;
        else Q22C = 0;
    if Q23 > 0 then Q23C = 1;
        else Q23C = 0;
    if Q24 > 0 then Q24C = 1;
        else Q24C = 0;
    if Q25 > 0 then Q25C = 1;
        else Q25C = 0;
    if Q26 > 0 then Q26C = 1;
        else Q26C = 0;
    if Q27 > 0 then Q27C = 1;
        else Q27C = 0;
    if Q28 > 0 then Q28C = 1;
        else Q28C = 0;
    attrib Q3aC Q3bC Q3cC Q3dC Q3eC format = Q3Cfmt.;
    attrib Q10C format = Q10Cfmt.;
    attrib Q17C Q18C Q19C Q20C Q21C format = SectCfmt.;
    attrib Q22C Q23C Q24C Q25C Q26C Q27C Q28C format = PartIIfmt.;

*************************************************************************************;
* Create cross tabs for questions with collapses categories                         ;
*************************************************************************************;

proc freq data=collapse;
    table Q3aC*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q3bC*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q3cC*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q3dC*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q3eC*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q10C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q17C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q18C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q19C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q20C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q21C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q22C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q23C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q24C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q25C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q26C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
    table Q27C*neurop / chisq exact norow nopercent;

proc freq data=collapse;
```

```
    table Q28C*neurop / chisq exact norow nopercent;

run;
```

### 9.6.4. Non-Collapsed Cross Tabulation

```
*******************************************************************************;
* Date:                7-22-97                                                 ;
* Last modified:       7-22-97                                                 ;
* Programmer:          John O'Gorman                                           ;
* Project:             Posture-Cross Validation                                ;
* Purpose:             Crosstab, chi-square analysis                           ;
* Program name:          Cross Tabulation Analysis.sas                                    ;
;
* Program location:    F:\Research\CELOS-Posture Cross Validation\Examples\    ;
* Data location:       F:\Research\CELOS-Posture Cross Validation\Examples\LogisticReg.csv;
* Output file:         LogisticReg.lst                                         ;
* Reference:           Chapter 16 of SAS/STAT Software Changes and Enhancements ;
*                      through Release 6.11                                    ;
*******************************************************************************;

options nodate nonumber linesize=64 pagesize=60;

* To read an Excel file into SAS, first delete out the column names and save it as  ;
* a CSV (comma delimited) file.  The line below tells SAS the location of the data file;
* and give it an internal name called in                                          ;

filename in 'F:\Scc Research\CELOS-Posture Cross Validation\Data\AnswersOnly.csv';

*******************************************************************************;
* Formats for categorical variable;
*******************************************************************************;

proc format;
    value neurfmt  0 = 'Non Neur'
                   1 = 'Neur';
    value yesnofmt 0 = 'No'
                   1 = 'Yes';
    value sectCfmt 0 = 'Very safe'
                   1 = 'Safe'
                   2 = 'Little unsafe'
                   3 = 'Very unsafe';


*******************************************************************************;
* Read in the AnswersOnly.csv data file  .                                      ;
*******************************************************************************;

data answers;
*   The name in refers to the name given in the filename statement above;
*   The delimiter=',' tells SAS that observations on each line are separated by a comma;
    infile in delimiter=',';
    input subject Q1 Q2 Q3a Q3b Q3c Q3d Q3e Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18
Q19 Q20 Q21
          Q22 Q22a Q23 Q23a Q24 Q24a Q25 Q25a Q26 Q26a Q27 Q27a Q28 Q28a group;
*   Drop out the non diabetics;
    if group = 1 then delete;
*   Define the neuropathy variable;
    if group > 2 then neurop = 1;
        else neurop = 0;
*   Delete subjects 11 and 48 who did not fill out the questionnaire;
    if subject = 11 or subject = 48 then delete;
*   Assigns the neurop variable the format described in proc format;
    attrib neurop label = 'Neuropathy' format = neurfmt.;
    attrib Q4 format = yesnofmt.;
    attrib Q17 format = sectCfmt.;

*   Drop out questions with missing values;
data complete;
    set answers (drop = Q10 Q18);
```

```
proc freq data=complete;
    table Q1*neurop / chisq exact;

proc freq data=complete;
    table Q2*neurop / chisq exact;

proc freq data=complete;
    table Q3a*neurop / chisq exact;

proc freq data=complete;
    table Q3b*neurop / chisq exact;

proc freq data=complete;
    table Q3c*neurop / chisq exact;

proc freq data=complete;
    table Q3d*neurop / chisq exact;

proc freq data=complete;
    table Q3e*neurop / chisq exact;

proc freq data=complete;
    table Q4*neurop / chisq exact;

proc freq data=complete;
    table Q5*neurop / chisq exact;

proc freq data=complete;
    table Q6*neurop / chisq exact;

proc freq data=complete;
    table Q7*neurop / chisq exact;

proc freq data=complete;
    table Q8*neurop / chisq exact;

proc freq data=complete;
    table Q9*neurop / chisq exact;

proc freq data=complete;
    table Q11*neurop / chisq exact;

proc freq data=complete;
    table Q12*neurop / chisq exact;

proc freq data=complete;
    table Q13*neurop / chisq exact;

proc freq data=complete;
    table Q14*neurop / chisq exact;

proc freq data=complete;
    table Q15*neurop / chisq exact;

proc freq data=complete;
    table Q16*neurop / chisq exact;

proc freq data=complete;
    table Q17*neurop / chisq exact;

proc freq data=complete;
    table Q19*neurop / chisq exact;

proc freq data=complete;
    table Q20*neurop / chisq exact;

proc freq data=complete;
    table Q21*neurop / chisq exact;

proc freq data=complete;
    table Q22*neurop / chisq exact;

proc freq data=complete;
    table Q23*neurop / chisq exact;

proc freq data=complete;
```

```
    table Q24*neurop / chisq exact;

proc freq data=complete;
    table Q25*neurop / chisq exact;

proc freq data=complete;
    table Q26*neurop / chisq exact;

proc freq data=complete;
    table Q27*neurop / chisq exact;

proc freq data=complete;
    table Q28*neurop / chisq exact;

proc freq data=complete;
    table Q3b*Q3c / measures;

run;
```

# 9.7. SAS Output

```
Logistic Regression

                    The LOGISTIC Procedure

Data Set: WORK.COLLAPSE
Response Variable: NEUROP
Response Levels: 2
Number of Observations: 44
Link Function: Logit


                      Response Profile

                  Ordered
                    Value  NEUROP     Count

                        1       1        28
                        2       0        16


            Model Fitting Information and Testing
                Global Null Hypothesis BETA=0

                          Intercept
           Intercept          and
Criterion    Only     Covariates Chi-Square for Covariates

AIC           59.682      48.434      .
SC            61.467      53.787      .
-2 LOG L      57.682      42.434    15.248 with 2 DF (p=0.0005)
Score            .           .      13.770 with 2 DF (p=0.0010)


            Analysis of Maximum Likelihood Estimates

                    Parameter   Standard      Wald         Pr >
 Variable   DF      Estimate      Error    Chi-Square   Chi-Square

 INTERCPT   1       -1.4122      0.8037       3.0879       0.0789
 Q22B       1        2.3393      0.8776       7.1052       0.0077
 Q24B       1        1.5206      0.8880       2.9322       0.0868

        Analysis of Maximum
       Likelihood Estimates

            Standardized        Odds
 Variable     Estimate          Ratio

 INTERCPT            .            .
 Q22B         0.649623       10.374
```

Q24B 0.355388 4.575

Logistic Regression

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

| | |
|---|---|
| Concordant = 70.8% | Somers' D = 0.625 |
| Discordant = 8.3% | Gamma = 0.791 |
| Tied = 21.0% | Tau-a = 0.296 |
| (448 pairs) | c = 0.813 |

Hosmer and Lemeshow Goodness-of-Fit Test

| | | NEUROP = 1 | |
|---|---|---|---|
| Group | Total | Observed | Expected |
| 1 | 8 | 2 | 1.57 |
| 2 | 16 | 8 | 8.43 |
| 3 | 2 | 1 | 1.43 |
| 4 | 18 | 17 | 16.57 |

Hosmer and Lemeshow Goodness-of-Fit Test

| NEUROP = 0 | |
|---|---|
| Observed | Expected |
| 6 | 6.43 |
| 8 | 7.57 |
| 1 | 0.57 |
| 1 | 1.43 |

Goodness-of-fit Statistic = 0.7993 with 2 DF (p=0.6706)

Classification Table

| | Correct | | Incorrect | | Percentages | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prob Level | Event | Non-Event | Event | Non-Event | Correct | Sensitivity | Specificity | False POS | False NEG |
| 0.120 | 28 | 0 | 16 | 0 | 63.6 | 100.0 | 0.0 | 36.4 | . |
| 0.140 | 26 | 0 | 16 | 2 | 59.1 | 92.9 | 0.0 | 38.1 | 100.0 |
| 0.160 | 26 | 0 | 16 | 2 | 59.1 | 92.9 | 0.0 | 38.1 | 100.0 |
| 0.180 | 26 | 0 | 16 | 2 | 59.1 | 92.9 | 0.0 | 38.1 | 100.0 |
| 0.200 | 26 | 0 | 16 | 2 | 59.1 | 92.9 | 0.0 | 38.1 | 100.0 |
| 0.220 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.240 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.260 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.280 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.300 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.320 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.340 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.360 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.380 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.400 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.420 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.440 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.460 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.480 | 26 | 6 | 10 | 2 | 72.7 | 92.9 | 37.5 | 27.8 | 25.0 |
| 0.500 | 18 | 6 | 10 | 10 | 54.5 | 64.3 | 37.5 | 35.7 | 62.5 |
| 0.520 | 18 | 6 | 10 | 10 | 54.5 | 64.3 | 37.5 | 35.7 | 62.5 |
| 0.540 | 18 | 6 | 10 | 10 | 54.5 | 64.3 | 37.5 | 35.7 | 62.5 |
| 0.560 | 18 | 14 | 2 | 10 | 72.7 | 64.3 | 87.5 | 10.0 | 41.7 |
| 0.580 | 18 | 14 | 2 | 10 | 72.7 | 64.3 | 87.5 | 10.0 | 41.7 |
| 0.600 | 18 | 14 | 2 | 10 | 72.7 | 64.3 | 87.5 | 10.0 | 41.7 |
| 0.620 | 18 | 14 | 2 | 10 | 72.7 | 64.3 | 87.5 | 10.0 | 41.7 |
| 0.640 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.660 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.680 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.700 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.720 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.740 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.760 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.780 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.800 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.820 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.840 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.860 | 17 | 14 | 2 | 11 | 70.5 | 60.7 | 87.5 | 10.5 | 44.0 |
| 0.880 | 17 | 15 | 1 | 11 | 72.7 | 60.7 | 93.8 | 5.6 | 42.3 |
| 0.900 | 17 | 15 | 1 | 11 | 72.7 | 60.7 | 93.8 | 5.6 | 42.3 |
| 0.920 | 0 | 15 | 1 | 28 | 34.1 | 0.0 | 93.8 | 100.0 | 65.1 |
| 0.940 | 0 | 15 | 1 | 28 | 34.1 | 0.0 | 93.8 | 100.0 | 65.1 |
| 0.960 | 0 | 16 | 0 | 28 | 36.4 | 0.0 | 100.0 | . | 63.6 |

Contents of rocdata data set

CONTENTS PROCEDURE

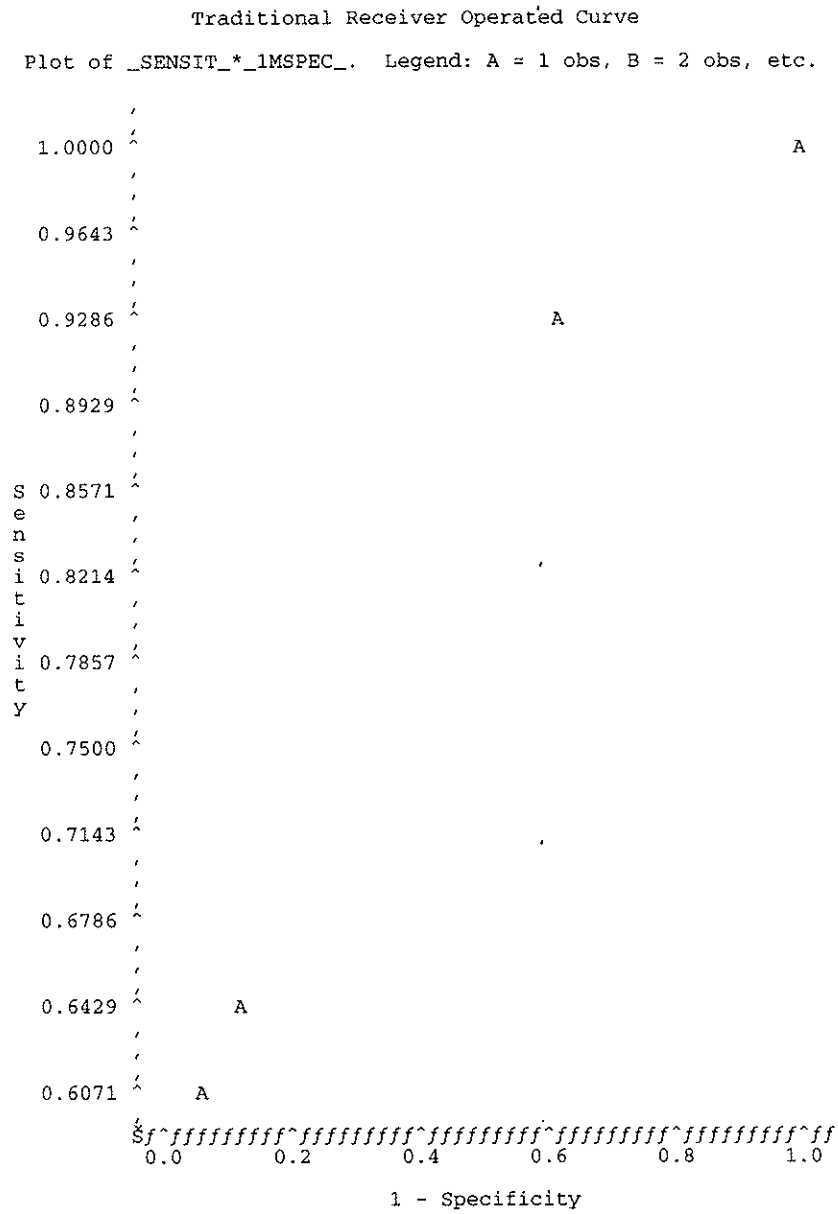
| | | | |
|---|---|---|---|
| Data Set Name: | WORK.ROCDATA | Observations: | 4 |
| Member Type: | DATA | Variables: | 7 |
| Engine: | V612 | Indexes: | 0 |
| Created: | 0:25 Thu, Aug 7, 1997 | Observation Length: | 56 |
| Last Modified: | 0:25 Thu, Aug 7, 1997 | Deleted Observations: | 0 |
| Protection: | | Compressed: | NO |
| Data Set Type: | | Sorted: | NO |
| Label: | Receiver Operating Characteristics | | |

-----Engine/Host Dependent Information-----

| | |
|---|---|
| Data Set Page Size: | 8192 |
| Number of Data Set Pages: | 1 |
| File Format: | 607 |
| First Data Page: | 1 |
| Max Obs per Page: | 145 |
| Obs in First Data Page: | 4 |

-----Alphabetic List of Variables and Attributes-----

| # | Variable | Type | Len | Pos | Label |
|---|---|---|---|---|---|
| 7 | _1MSPEC_ | Num | 8 | 48 | 1 - Specificity |
| 5 | _FALNEG_ | Num | 8 | 32 | No. of Events Predicted as Nonevents |
| 4 | _FALPOS_ | Num | 8 | 24 | No. of Nonevents Predicted as Events |
| 3 | _NEG_ | Num | 8 | 16 | No. of Correctly Predicted Nonevents |
| 2 | _POS_ | Num | 8 | 8 | No. of Correctly Predicted Events |
| 1 | _PROB_ | Num | 8 | 0 | Probability Level |
| 6 | _SENSIT_ | Num | 8 | 40 | Sensitivity |


Traditional Receiver Operated Curve
Plot of _SENSIT_*_1MSPEC_. Legend: A = 1 obs, B = 2 obs, etc.
1.0000
0.9643
0.9286
0.8929
0.8571
0.8214
0.7857
0.7500
0.7143
0.6786
0.6429
0.6071
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity