Optimal Data Analysis
2013, Vol. 2, Release 1 (September 27, 2013), 43-47

2155-0182/11/$3.00

Reverse CTA Versus Multiple Regression Analysis

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

This paper illustrates how to reverse CTA for applications having an ordered class variable and categorical attributes. Whereas a regression model is used to make point predictions for the dependent measure based on values of the independent variables, reverse CTA is used to find domains on the dependent measure which are explained by the independent variables.

*S*elf-*m*onitoring *a*nd *r*eview *t*ool (SMART) is an interactive, internet-based, self-monitoring and feedback system for helping individuals identify and monitor the relationships between their own behaviors, stressors, management strategies and symptom levels across time.[1] SMART involves longitudinal collection and statistical analysis of self-monitoring data, with the ultimate objective being the timely delivery of personalized feedback derived from the data.[2]

The present study examines longitudinal data for an individual using SMART to rate the intensity of nine fibromyalgia (FM) symptoms experienced over 297 consecutive days. Rated using 10-point Likert-type response scales, the symptoms are pain, stiffness, fatigue, concentration problems, memory problems, anxiety, depression, gastrointestinal problems, and sleep problems.[1] These scales serve as independent variables in multiple regression analysis (MRA), or analogously as attributes in reverse CTA.

Presently, the dependent (MRA) or class (reverse CTA) variable is 500 mb geopotential height anomaly (HT500) measured in meters, an atmospheric pressure index independently recorded by the investigator (RCS) for every day in the longitudinal record.[2] Descriptive statistics for study variables are provided in Table 1.

Multiple Regression Analysis

Among the most widely used statistical analysis methods, MRA requires little in the way of introduction.[3]

The present data were analyzed by MRA for expository purposes. The first analysis used raw data, HT500 as the dependent variable, and symptom ratings as the independent variables. Using all nine symptoms the model had R^2=0.34 [F(9,287)=16.4, p<0.0001], with concentration problems (p<0.0001), fatigue (p<0.0001), and anxiety (p<0.02) explaining statistically reliable unique variance. Analysis using only these three independent variables found the effect for anxiety was statistically unreliable, so the final model used concentration problems and fatigue (p's<0.0001) as independent variables: R^2=0.23; F(2,294)=43.1, p<0.0001. The regression model relating HT500 to patient symptoms was:

HT500=5716.6+36.0*concentration problems-68.8*fatigue.

2155-0182/11/$3.00

Table 1: Study Variable Descriptive Statistics

Variable	Mean	SD	Median
HT500	5575	158	5565
Pain	4.6	1.5	5
Stiffness	5.3	1.6	5
Fatigue	6.4	1.4	6
Concentration	5.4	1.7	5
Memory	5.6	1.4	6
Anxiety	1.0	0.9	1
Depression	1.5	1.6	1
Gastrointestinal	0.4	1.0	0
Sleep	5.6	1.6	5

Note: HT500 is an index of atmospheric pressure which is assessed in meters. The remaining variables are symptom ratings made using 10-point (0-9) Likert-type scales on which increasing values indicate worsening symptoms. SD=standard deviation. For reverse CTA a daily symptom was coded as positive if it exceeded the median value, and as negative otherwise.

The model shows that for this person, increasing HT500 is associated with decreasing fatigue and increasing concentration problems.

In the second analysis dummy-variable MRA was performed with median-based binary symptom indicators as independent variables (concentration problems was dropped to prevent multicollinearity). With all eight symptoms the model had R^2=0.17 [F(8,288)=7.5, p<0.0001], with stiffness (p<0.0001) and anxiety (p<0.002) explaining statistically reliable unique variance. The final model relating HT500 to symptoms using stiffness (p<0.0001) and anxiety (p<0.04) as the independent variables [R^2=0.13, F(2,294) =22.8, p<0.0001] was:

HT500=5618.2–110.6*stiffness+
49.2*anxiety.

The model shows that for this person, increasing HT500 is associated with decreasing stiffness and increasing anxiety.

Motivating Reverse CTA

A challenging pragmatic issue for the regression results, in light of the study aim, is how the regression models may be used to alert patients about forthcoming symptom-inducing or symptom-relieving weather. For example, if a forecast calls for higher HT500, what does this imply about levels of stiffness and anxiety the patient should anticipate experiencing?

While the regression equation may be rearranged to estimate one symptom based on HT500, it is impossible to estimate both of the symptoms simultaneously. By their structure it is obvious that the regression models are useful for predicting the level of HT500 from patient symptoms. Instead what is needed is precisely the opposite functionality, predicting the patient symptoms based on HT500. As demonstrated below, reverse CTA provides this functionality.

Defining Attributes

In conventional CTA the class variable is binary, and attributes may be categorical or ordered.[4] In reverse CTA the class variable is ordered, and *attributes must be categorical*. The present research simulates alerts about relatively severe symptom days, defined for each day and symptom as positive if the rating is greater than the median value for the symptom (see Table 1), and as negative otherwise.

Reverse CTA begins by determining the attributes to include in analysis, using structural decomposition.[5] In step one the strength of the relationship between HT500 and each of the nine categorical variables was evaluated for the total sample, and the model for stiffness had greatest associated ESS. The UniODA model was: if HT500≤5675 ppm predict positive, otherwise predict negative. This model achieved moderate ESS=40.0: it correctly classified 60% of N=214 days having HT500 below the cut-point, and 89% of N=83 days having HT500 above the cut-point (p<0.0001).

2155-0182/11/$3.00

In step two depression had greatest ESS (41.4, $p<0.0001$), and the UniODA model (if HT500≤5490 ppm then predict positive) correctly classified 60% of N=40 days with HT500 below the cut-point, and 80% of N=54 days with HT500 above the cut-point.

In the third and final step anxiety had greatest ESS (75.0, $p<0.05$), and the UniODA model (if HT500>5560 ppm then predict positive) correctly classified 100% of N=18 days having HT500 below the cut-point, and 33.3% of N=9 days having HT500 above the cut-point. Thus, stiffness, depression and anxiety were selected as the attributes to use in reverse CTA.

In conventional analysis these attributes would be analyzed with CTA conducted using automated software.[4]

Obtaining the Model

Unfortunately no automated software is available which is capable of performing reverse CTA, so the analysis is conducted manually vis-à-vis UniODA software.[5] Because only a small number of attributes were identified presently in structural decomposition analysis, enumerated reverse CTA is demonstrated. Decomposition analysis determined that the model for stiffness had strongest ESS, and it is arbitrarily selected as the root attribute in the first of the three CTA models required to conduct the enumeration. Decomposition and reverse CTA analysis are both performed using the following UniODA software[5] code (25,000 monte carlo experiments were run to obtain the Type I error[6]; software control commands are indicated in red:

```
open example.dat;
output example.out;
vars ht500 stiff depress anxiety;
class ht500;
attr stiff depress anxiety;
mc iter 25000;
go;
```

An illustration of the root of this reverse CTA model is provided in Figure 1. Ordinarily attributes are shown in nodes and class variable in endpoints, however the opposite order occurs in reverse CTA. Similarly to conventional CTA, arrows indicate paths from class variables to attributes: however in reverse CTA arrows point *up* the tree. In contrast to conventional CTA where one reads down the tree starting from the root, in reverse CTA one reads up the tree starting from the endpoints. As seen, when HT500≤ 5675m, stiffness is correctly predicted to be positive on 129 of 214 (60.3%) days, and when HT500>5675m, stiffness is correctly predicted to be negative on 74 of 83 (89.2%) days.

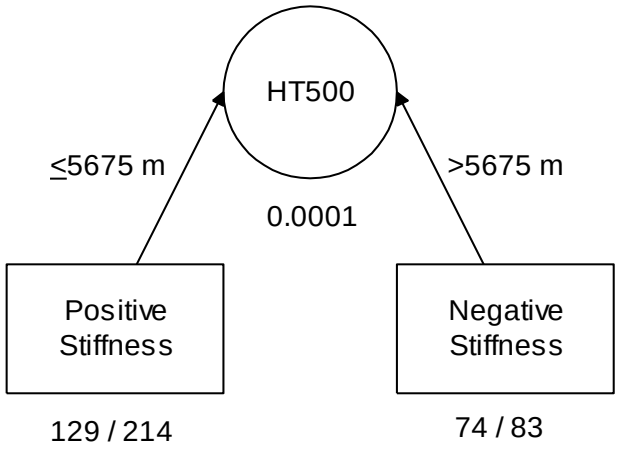

Figure 1: Root of First Reverse CTA Model

The next analytic step involves assessing whether an attribute should be added to the left endpoint, requiring *adding* the command:

```
exclude ht500>5675;
```

Depression had the greatest ESS (28.4, $p<0.0003$) and was thus added to the model, as illustrated in Figure 2. No additional variables could be used as left (p's>0.14) or right (p's> 0.25) branches off depression, so construction of the left-hand side of the model is complete.

2155-0182/11/$3.00

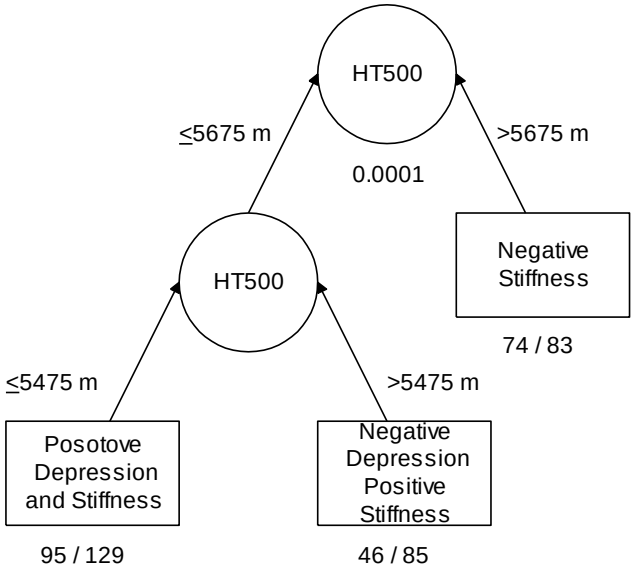

Figure 2: Attribute Added on Left Branch

The next step involves assessing whether to add an attribute to the right endpoint. This is accomplished by *modifying* the prior command:

exclude ht500<=5675;

The UniODA model for anxiety had the greatest ESS and was added to the model. The classification for the left endpoint was perfect so no attribute could be added, and no statistically significant attributes emerged at the right-hand branch: thus, construction of this reverse CTA model is complete (see Figure 3).

The next step involves pruning the full CTA model in order to find the (sub)model with greatest overall ESS.[7] The reverse CTA model in Figure 2 had greatest overall ESS of 44.5 (versus 30.4 for the full model in Figure 3).

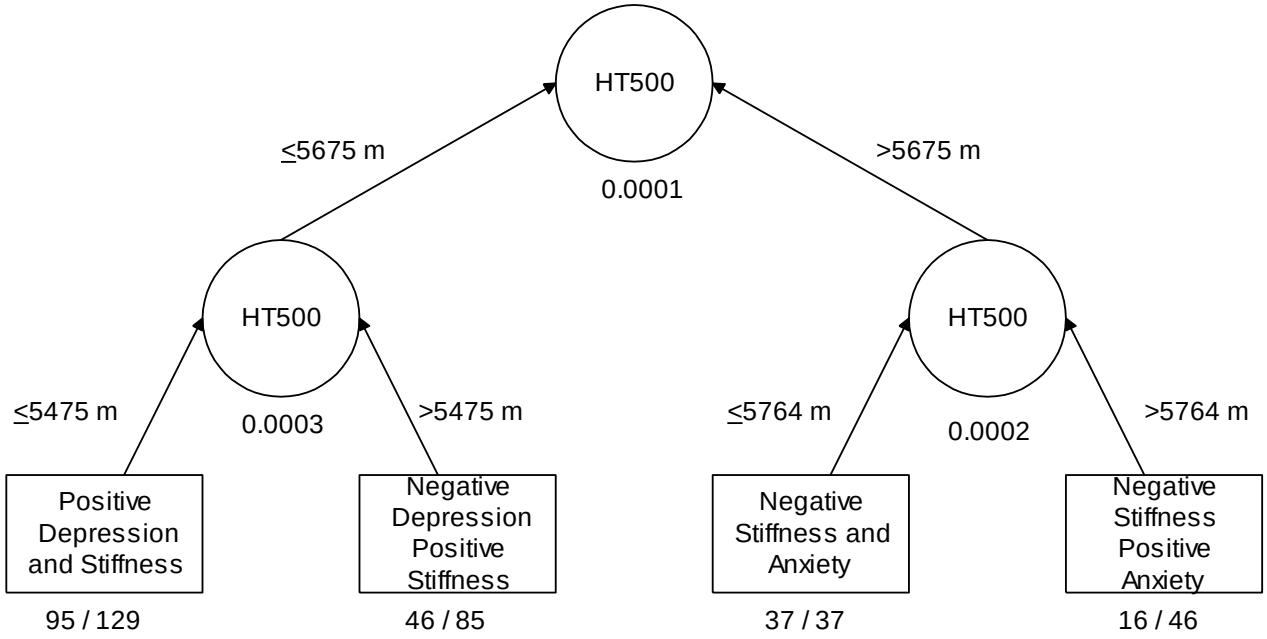

Figure 3: Complete Non-Pruned Reverse CTA Model Having Stiffness as the Root Variable

The foregoing procedure is now repeated twice, once using depression as root, and once again using anxiety as root. Not illustrated, ESS of the resulting pruned reverse CTA models was 39.3 and 22.7, respectively.

Discussion

Thus, the reverse CTA model illustrated in Figure 2 was the strongest (ESS=44.5), most efficient (ESS=22.2 per node), and also the most parsimonious representation of the longitudinal

2155-0182/11/$3.00

symptom data identified. As seen in Table 2, interpretation of the model can be simplified by examining the corresponding staging table.

Table 2: Staging Table Predicting Patient Symptoms as a Function of HT100 Values

Stage	HT500	Symptom	N	%	Odds
1	>5675 m	None	83	89.2	1:8
2	>5475 m	Stiffness	85	53.4	1:1
3	≤5475 m	Stiffness, Depression	129	76.0	3:1

Note: HT500 is a measure of atmospheric pressure. N is the number of days in the series with HT500 falling in the indicated domain (stage), and % is the percent of the N days in which the indicated symptom was in fact present. Odds are given for a *bad symptom day*.

In the case of the present individual, the odds of a bad symptom day are 1 in 9 when the pressure is high (HT500>5675m); 1 in 2 (for stiffness) when pressure falls to an intermediate level (HT500>5475m); and 3 in 4 when the pressure falls to a low level (HT500≤5475m).

Several weather services predict HT500 days to two weeks or longer into the future. Thus, providing individuals with short-term and intermediate-range alerts regarding the odds of experiencing good- and bad-symptom days is a realistic opportunity.

References

[1]Collinge W, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis, 1*, 163-175.

[2]Yarnold PR, Soltysik RC, Collinge W (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis, 2*, 43-48.

[3]Licht MH (1995). Multiple regression and correlation. In Grimm LG, Yarnold PR. (Eds.), *Reading and understanding multivariate statistics*. APA Books, pp. 19-64.

[4]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*: 144-160.

[5]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, APA Books.

[6]Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo estimation of two-category UniODA two-tailed *p*. *Optimal Data Analysis*, *1*: 43-45.

[7]Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis*, *1*, 23-29.

Author Notes

Correspondence by mail should be sent to: Optimal Data Analysis LLC, 1220 Rosecrans St., #330, San Diego, CA 92106. Send e-mail to: Journal@OptimalDataAnalysis.com.