Ordinal Independent Variables
Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised April 9, 2017

Long & Freese, 2006, Regression Models for Categorical Dependent Variables Using Stata, Second Edition (Not the third!)
Many of the ideas presented here were also discussed in http://www.statalist.org/forums/forum/general-stata-
discussion/general/1330543-annoyingly-coded-ordinal-independent-variables.

We often want to use ordinal variables as independent/explanatory variables in our models. Rightly or wrongly, it is very common to treat such variables as continuous. Or, more precisely, as having interval-level measurement with linear effects. When the items uses a Likert scale (e.g. Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) this may be a reasonable practice. However, many ordinal items use categories that clearly are not equally spaced, e.g. the options might be something like “daily,” “a few times a week,” “once a week”, “a few times a month”,... “once a year”, “never.”

In the paper referenced above, David J. Pasta makes a strong case for (usually) treating ordinal variables as continuous, even when the spacing is not equal across categories. He says (pp. 2 -3)

One concern often expressed is that “we don't know that the ordinal categories are equally spaced.” That is true enough – we don't. But we also don't “know” that the relationship between continuous variables is linear, which means we don't “know” that a one-unit change in a continuous variable has the same effect no matter whether it is a change between two relatively low values or a change between two relatively high values. In fact, when it's phrased that way -- rather than “is the relationship linear?” -- I find a lot more uncertainty in my colleagues. It turns out that it doesn't matter that much in practice -- the results are remarkably insensitive to the spacing of an ordinal variable except in the most extreme cases. It does, however, matter more when you consider the products of ordinal variables.

I am squarely in the camp that says “everything is linear to a first approximation” and therefore I am very cheerful about treating ordinal variables as continuous. Deviations from linearity can be important and should be considered once you have the basics of the model established, but it is very rare for an ordinal variable to be an important predictor and have it not be important when considered as a continuous variable. That would mean that the linear component of the relationship is negligible but the non-linear component is substantial. It is easy to create artificial examples of this situation, but they are very, very rare in practice.

To elaborate on one of Pasta’s points – Even variables with interval-level coding don't necessarily have linear effects. You may need to take logs, add squared terms, estimate spline functions, etc. I think the issue is just a bit more obvious with ordinal variables because the number of possible values is limited and it is often questionable to believe that the categories are equally spaced.

Long and Freese (in the 2006 edition of their book) agree that ordinal variables are often treated as continuous. But they add (p. 421) that
The advantage of this approach is that interpretation is simpler, but to take advantage of this simplicity you must make the strong assumption that successive categories of the ordinal independent variable are equally spaced. For example, it implies that an increase from no publications by the mentor to a few publications involves an increase of the same amount of productivity as an increase from a few to some, from some to many, and from many to lots of publications. Accordingly, before treating an ordinal independent variable as if it were interval, you should test whether this leads to a loss of formation about the association between the independent and dependent variable.

In short, it will often be ok to treat an ordinal variable as though it had linear effects. The greater parsimony that results from doing so may be enough to offset any disadvantages that result. But, there are ways to formally test whether the assumption of linearity is justified.

**Likelihood Ratio Chi-Square Test and/or BIC tests.** Here, you estimate two models. In the constrained model the ordinal variable is treated as continuous, in the unconstrained model it is treated as categorical. You then use an LR chi-square test (or a BIC test or AIC test) to decide whether use of the more parsimonious continuous measure is justified.

```
webuse nhanes2f, clear
* Treat health as continuous
. logit diabetes c.health, nolog
Logistic regression                           Number of obs     =     10,335
LR chi2(1)        =     428.14
Prob > chi2       =     0.0000
Log likelihood = -1784.9973                     Pseudo R2         =     0.1071
------------------------------------------------------------------------------
diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    health |  -.8128706   .0421577   -19.28   0.000    -.8954982    -.730243
      _cons |   -.637503   .1113932    -5.72   0.000    -.8558296   -.4191764
------------------------------------------------------------------------------
. est store m1

* Now treat health as categorical
. logit diabetes i.health, nolog
Logistic regression                           Number of obs     =     10,335
LR chi2(4)        =     429.74
Prob > chi2       =     0.0000
Log likelihood = -1784.1984                     Pseudo R2         =     0.1075
------------------------------------------------------------------------------
diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     health |  -.7493387   .0421577   -17.80   0.000    -.831566     -.667111
    average |  -1.567205   .1302544   -12.03   0.000    -1.822499   -1.311911
       good |  -2.554012   .1780615   -14.34   0.000    -2.903006   -2.205018
  excellent |  -3.116457   .2262238   -13.78   0.000    -3.559848   -2.673067
      _cons |  -1.4516605   .0953463   -15.14   0.000    -1.636816   -1.266508
------------------------------------------------------------------------------
. est store m2
```
. * Now do LR/ BIC/ AIC tests
. lrtest m1 m2, stats

Likelihood-ratio test
(Assumption: m1 nested in m2)

LR chi2(3) = 1.60
Prob > chi2 = 0.6599

A visual inspection of the coefficients from the 2nd model indeed suggests that the effects of health are continuous, i.e. each coefficient is about .75 greater than the coefficient before it. The LR/ BIC/ AIC tests also all agree that the more parsimonious model that treats health as a continuous variable is preferable.

Wald tests. Of course, you can’t always do LR tests. Luckily, Wald tests are also possible. One way to do this is by including both the continuous and categorical versions of the ordinal variable in the analysis. If the effects of the categorical variable are not statistically significant, then the continuous version alone is sufficient. Note that, because we are including two versions of the ordinal variable, two categories of the ordinal variable must be excluded rather than the usual one. We can do this via use of the o. notation (o stands for omitted).

. * Wald test
. logit diabetes c.health o(1 2).health, nolog

Logistic regression
Number of obs = 10,335
LR chi2(4) = 429.74
Prob > chi2 = 0.0000
Log likelihood = -1784.1984
Pseudo R2 = 0.1075

| diabetes | Coef. | Std. Err. | z  | P>|z| | [95% Conf. Interval] |
|-----------|-------|-----------|----|-----|---------------------|
| health    | -.7493387 | .1262017 | -5.94 | 0.000 | -.9966895, -.5019878 |
| health    |        |           |     |  | fair (omitted) |
| average   | -.0685278 | .2104996 | -0.33 | 0.745 | -.4810995, .3440439 |
| good      | -.3059957 | .3471345 | -0.88 | 0.378 | -.9863668, .3743754 |
| excellent | -.1191024 | .4829907 | -0.25 | 0.805 | -1.065747, .8275419 |
| _cons     | -.7322659 | .2078451 | -3.52 | 0.000 | -1.139635, -.3248969 |

. testparm i.health

( 1) [diabetes]3.health = 0
( 2) [diabetes]4.health = 0
( 3) [diabetes]5.health = 0

   chi2(  3) = 1.56
Prob > chi2 = 0.6689
Again, the results indicate that the continuous version of the variable is fine.

Other options. Other strategies for dealing with ordinal independent variables have been proposed. In the Statalist thread linked to above, Ben Earnhart notes that a common, if not necessarily correct approach, is to code the midpoint of categories, e.g.

“daily”=1
“once a week”=1/7
“a few times a month”=(1/7)*(3/4)
“once a year”=1/365.25
“never”=0

Of course, some categories may be open-ended (e.g. $100,000 or more) which can make this strategy problematic.

Maarten Buis (same thread) also suggests that the `sheafcoef` command (available from SSC) can be used. The help file for `sheafcoef` says

```
sheafcoef is a post-estimation command that estimates sheaf coefficients (Heise 1972). A sheaf coefficient assumes that a block of variables influence the dependent variable through a latent variable. sheafcoef displays the effect of the latent variable and the effect of the observed variables on the latent variable. The scale of the latent variable is identified by setting the standard deviation equal to one. The origin of the latent variable is identified by setting it to zero when all observed variables in its block are equal to zero. This means that the mean of the latent variable is not (necessarily) equal to zero. The final identifying assumption is that the effect of the latent variable is always positive, so to give a substantive interpretation of the direction of the effect, one needs to look at the effects of the observed variables on the latent variable. Alternatively, one can specify one “key” variable in each block of variables, which identifies the direction of a latent variable, either by specifying that the latent variable has a high value when the key variable has a high value or that the latent variable has a low value when the key variable has a high value.

The assumption that the effect of a block of variables occurs through a latent variable is not a testable constraint; it is just a different way of presenting the results from the original model. Its main usefulness is in comparing the relative strength of the influence of several blocks of variables. For example, say we want to know what determines the probability of working non-standard hours and we have a block of variables representing characteristics of the job and another block of variables representing the family situation of the respondent, and we want to say something about the relative importance of job characteristics versus family situation. In that case one could estimate a logit model with both blocks of variables and optionally some other control variables. After that one can use sheafcoef to display the effects of two latent variables, family background and job characteristics, which are both standardized to have a standard deviation of 1, and can thus be more easily compared.

The output is divided into a number of equations. The top equation, labeled “main”, represents the effects of the latent variables and other control variables (if any) on the dependent variable. The names of the latent variables are as specified in the latent() option. If no names are specified, they will be called “ivar1”, “ivar2”, etc. Below the main equation, one additional equation for every latent variable is displayed, labelled “on_name1”, “on_name2”, etc., where “name1” and “name2” are the names of the latent variables. These are the effects of the observed variables on the latent variable.
```
The sheaf coefficients and the variance covariance matrix of all the coefficients are estimated using `nlcom`. `sheafcoef` can be used after any regular estimation command (that is, a command that leaves its results behind in e(b) and e(V)). The only constraint is that the observed variables that make up the latent variable(s) must all come from the same equation.

Here is an example. As far as I know, `sheafcoef` does not support factor variables, so we have to compute the dummies ourselves.

```stata
* sheaf coefficients
.tab health, gen(hlth)

. logit diabetes hlth2 hlth3 hlth4 hlth5, nolog
Logistic regression                             Number of obs     =     10,335
LR chi2(4)        =     429.74
Prob > chi2       =     0.0000
Log likelihood = -1784.1984                     Pseudo R2         =     0.1075
------------------------------------------------------------------------------
diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hlth2 |  -.7493387   .1262017    -5.94   0.000    -.9966895   -.5019878
    hlth3 |  -1.567205   .1302544   -12.03   0.000    -1.822499   -1.311911
    hlth4 |  -2.554012   .1780615   -14.34   0.000    -2.903006   -2.205018
    hlth5 |  -3.116457   .2262238   -13.78   0.000    -3.559848   -2.673067
      _cons |  -1.481605   .0953463   -15.54   0.000     -1.66848   -1.294729
------------------------------------------------------------------------------
.sheafcoef, latent(hlth: hlth2 hlth3 hlth4 hlth5)

------------------------------------------------------------------------------
diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     main    |
       hlth |   .9751702   .0668694    14.58   0.000     .8441085    1.106232
      _cons |  -1.481605   .0953463   -15.54   0.000    -1.66848   -1.294729
     on_hlth|
      hlth2 |  -.7684183   .1401325    -5.48   0.000    -1.043073   -.4937637
      hlth3 |  -1.607109   .1688974    -9.52   0.000    -1.938142   -1.276077
      hlth4 |  -2.619042   .1967101   -13.31   0.000    -3.004587   -2.233497
      hlth5 |  -3.195808   .1154059   -27.69   0.000     -3.422   -2.969617
------------------------------------------------------------------------------
```

In short, the main equation tells us how the underlying latent variable hlth affects the dependent variable diabetes. The on_hlth equation shows you how the observed hlth dummies affect the latent variable hlth. You don’t need to assume that the categories are equally spaced.
Here is another example. In this case the LR test says we should NOT treat the ordinal variable agegrp as continuous. Visual inspection of the coefficients in the model that treats agegrp as categorical also suggests that it may not be correct to treat the effects of the variable as linear. However the BIC test disagrees, so a reasonable case could be made for going with the more parsimonious model.

* Another example: agegrp

```
. logit diabetes c.agegrp, nolog
Logistic regression                              Number of obs     =     10,335
LR chi2(1)        =     326.98
Prob > chi2       =     0.0000
Log likelihood = -1835.5776                     Pseudo R2         =     0.0818
------------------------------------------------------------------------------
diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
agegrp |   .5533155   .0350691    15.78   0.000     .4845813    .6220497
  _cons |  -5.216729   .1683673   -30.98   0.000    -5.546723   -4.886735
------------------------------------------------------------------------------
```

```
est store m1
. logit diabetes i.agegrp, nolog
Logistic regression                              Number of obs     =     10,335
LR chi2(5)        =     337.17
Prob > chi2       =     0.0000
Log likelihood = -1830.4836                     Pseudo R2         =     0.0843
------------------------------------------------------------------------------
diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age30-39 |   .7021745   .3396247     2.07   0.039     .0365223    1.367827
age40-49 |  1.660128    .3028614     5.48   0.000      1.06653    2.253725
age50-59 |  2.207308    .2860264     7.72   0.000     1.646706    2.767909
age60-69 |  2.638420    .2677401     9.85   0.000     2.113659    3.163181
age 70+  |  2.971236    .2779455    10.69   0.000     2.426472    3.515999
  _cons |  -5.034786   .2590377   -19.44   0.000    -5.54249   -4.527081
------------------------------------------------------------------------------
```

```
est store m2
. lrtest m1 m2, stats
```

```
Likelihood-ratio test                                LR chi2(4)  =     10.19
(Assumption: m1 nested in m2)                        Prob > chi2 =    0.0374

Akaike's information criterion and Bayesian information criterion

<table>
<thead>
<tr>
<th>Model</th>
<th>Obs</th>
<th>ll(null)</th>
<th>ll(model)</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>10,335</td>
<td>-1999.067</td>
<td>-1835.578</td>
<td>2</td>
<td>3675.155</td>
<td>3689.642</td>
</tr>
<tr>
<td>m2</td>
<td>10,335</td>
<td>-1999.067</td>
<td>-1830.484</td>
<td>6</td>
<td>3672.967</td>
<td>3716.427</td>
</tr>
</tbody>
</table>
```

Note: N=Obs used in calculating BIC; see [R] BIC note.
Using `sheafcoef`,

* Sheaf coefficients for agegrp

```
. tab agegrp, gen(xage)
```

<table>
<thead>
<tr>
<th>Age groups</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>age20-29</td>
<td>2,320</td>
<td>22.44</td>
<td>22.44</td>
</tr>
<tr>
<td>age30-39</td>
<td>1,621</td>
<td>15.68</td>
<td>38.13</td>
</tr>
<tr>
<td>age40-49</td>
<td>1,270</td>
<td>12.29</td>
<td>50.41</td>
</tr>
<tr>
<td>age50-59</td>
<td>1,289</td>
<td>12.47</td>
<td>62.88</td>
</tr>
<tr>
<td>age60-69</td>
<td>2,852</td>
<td>27.59</td>
<td>90.47</td>
</tr>
<tr>
<td>age 70+</td>
<td>985</td>
<td>9.53</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>10,337</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

```
. logit diabetes xage2 xage3 xage4 xage5 xage6, nolog
```

Logistic regression  Number of obs = 10,335
LR chi2(5) = 337.17
Prob > chi2 = 0.0000
Log likelihood = -1830.4836  Pseudo R2 = 0.0843

|              | Coef.   | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|--------------|---------|-----------|-------|-----|---------------------|
| diabetes     |         |           |       |     |                     |
|              |         |           |       |     |                     |
| xage2        | .7021745| .3396247  | 2.07  | 0.039| .0365223 - 1.367827 |
| xage3        | 1.660128| .3028614  | 5.48  | 0.000| 1.06653 - 2.253725  |
| xage4        | 2.207308| .2860264  | 7.72  | 0.000| 1.646706 - 2.767909 |
| xage5        | 2.638420| .2874101  | 9.25  | 0.000| 2.113659 - 3.16318  |
| xage6        | 2.971236| .2779455  | 10.69 | 0.000| 2.426472 - 3.515999 |
| _cons        | -5.034786| .2590377  | -19.44| 0.000| -5.54249 - -4.527081|

```
. sheafcoef, latent(age: xage2 xage3 xage4 xage5 xage6)
```

| diabetes     | Coef.   | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|--------------|---------|-----------|-------|-----|---------------------|
|              |         |           |       |     |                     |
| main         |         |           |       |     |                     |
| age          | 1.106507| .0915181  | 12.09 | 0.000| .9271344 - 1.285879 |
| _cons        | -5.034786| .2590377  | -19.44| 0.000| -5.54249 - -4.527081|
| on_age       |         |           |       |     |                     |
| xage2        | .6345868| .2841502  | 2.23  | 0.026| .0776627 - 1.191511 |
| xage3        | 1.500333| .1910889  | 7.85  | 0.000| 1.125805 - 1.87486 |
| xage4        | 1.994844| .1405728  | 14.19 | 0.000| 1.719326 - 2.270362 |
| xage5        | 2.384459| .0891692  | 26.74 | 0.000| 2.209691 - 2.559227 |
| xage6        | 2.685240| .1076525  | 24.94 | 0.000| 2.474245 - 2.896235 |

In short, with `sheafcoef`, we potentially get the advantages of treating an ordinal variable as continuous, without actually having to assume that categories are equally spaced. Whether it is worth the trouble is another matter; you can judge based on the circumstances. It may depend on what the tests of linear effects say or how reasonable it is to treat a variable as continuous based on its coding.