

Standardized Coefficients in Logistic Regression

NOTE: Long and Freese's `spost9` programs are used in this handout; specifically, the `listcoef` command, which is part of `spost9`, is used. Use the `findit` command to locate and install `spost9`. See Long and Freese's book, Regression Models for Categorical Dependent Variables Using Stata, Second Edition, for more information. Long's 1997 Regression Models for Categorical and Limited Dependent Variables provides a brief substantive discussion on pp. 69-71.

Overview. Long and Freese discuss alternative ways of standardizing variables that may help with interpretation. They primarily talk about these techniques with regards to logistic, multinomial logistic, and ordinal regression models, but they may be useful for OLS regression as well. Their `listcoef` command illustrates these different alternatives. I'll first present some preliminary results that will make it easier to understand what `listcoef` is doing.

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-logit.dta, clear
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
grade	32	.34375	.4825587	0	1
gpa	32	3.117188	.4667128	2.06	4
tuce	32	21.9375	3.901509	12	29
psi	32	.4375	.5040161	0	1

```
. logit grade gpa tuce psi
```

```
Iteration 0: log likelihood = -20.59173
Iteration 1: log likelihood = -13.496795
Iteration 2: log likelihood = -12.929188
Iteration 3: log likelihood = -12.889941
Iteration 4: log likelihood = -12.889633
Iteration 5: log likelihood = -12.889633
```

```
Logit estimates                                Number of obs =          32
                                                LR chi2(3)           =          15.40
                                                Prob > chi2          =          0.0015
Log likelihood = -12.889633                    Pseudo R2            =          0.3740
```

grade	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gpa	2.826113	1.262941	2.24	0.025	.3507938	5.301432
tuce	.0951577	.1415542	0.67	0.501	-.1822835	.3725988
psi	2.378688	1.064564	2.23	0.025	.29218	4.465195
_cons	-13.02135	4.931325	-2.64	0.008	-22.68657	-3.35613

```
. fitstat
```

```
Measures of Fit for logit of grade
```

```
Log-Lik Intercept Only:      -20.592      Log-Lik Full Model:      -12.890
D(28):                       25.779      LR(3):                   15.404
                              Prob > LR:           0.002
McFadden's R2:               0.374      McFadden's Adj R2:      0.180
Maximum Likelihood R2:       0.382      Cragg & Uhler's R2:     0.528
McKelvey and Zavoina's R2:   0.544      Efron's R2:             0.426
Variance of y*:              7.210      Variance of error:      3.290
Count R2:                    0.813      Adj Count R2:           0.455
AIC:                          1.056      AIC*n:                  33.779
BIC:                          -71.261     BIC':                   -5.007
```

```
. listcoef, std help
```

```
logit (N=32): Unstandardized and Standardized Estimates
```

```
Observed SD: .4825587
Latent SD: 2.6850837
```

```
Odds of: 1 vs 0
```

```
-----+-----+-----+-----+-----+-----+-----+-----+
      grade |      b      z      P>|z|      bStdX      bStdY      bStdXY      SDofX
-----+-----+-----+-----+-----+-----+-----+-----+
      gpa |  2.82611  2.238  0.025  1.3190  1.0525  0.4912  0.4667
      tuce |  0.09516  0.672  0.501  0.3713  0.0354  0.1383  3.9015
      psi |  2.37869  2.234  0.025  1.1989  0.8859  0.4465  0.5040
-----+-----+-----+-----+-----+-----+-----+-----+

```

```
      b = raw coefficient
      z = z-score for test of b=0
      P>|z| = p-value for z-test
      bStdX = x-standardized coefficient
      bStdY = y-standardized coefficient
      bStdXY = fully standardized coefficient
      SDofX = standard deviation of X
```

In the `listcoef` output, the column labeled `b` (which the `logit` command labels as `Coef.`) gives the unstandardized (metric) coefficients. The columns labeled `z` and `P>|z|` are also the same as in the `logit` output. The other columns (which were presented because I used the `std` option) give information that is relevant to different types of standardization. The `help` option added the descriptions of what each part of the output means.

Full Standardization. With full standardization, both the `X` and the `Y*` variables are standardized to have a mean of 0 and a standard deviation of 1. It is similar to standardization in OLS regression (with the important difference that `Y*` is a latent variable and not observed; we'll see why this is important below). In the `listcoef` output, the fully standardized coefficients are in the column labeled `bStdXY`. [NOTE: As `fitstat` shows, the variance of `Y*` is 7.21, which means its standard deviation is 2.685 – the same as what `listcoef` reports.]

The results show you that a 1 standard deviation increase in `gpa` results, on average, in almost half a standard deviation increase (.4912) in the log odds of getting an A.

If you know the metric coefficients and the standard deviations of the the `x`'s and `y*`, you can compute the standardized coefficients the same way you do in OLS:

$$b'_k = b_k * \frac{S_{x_k}}{S_{y^*}}$$

So, for example, to get the fully standardized effect of gpa,

$$b'_{gpa} = b_{gpa} * \frac{S_{gpa}}{S_{y^*}} = 2.82611 * \frac{.4667}{2.685} = .4912$$

X-Standardization. An intermediate approach is to standardize only the X variables. In the `listcoef` output, in the column labeled `bStdX`, the Xs are standardized but Y* is not. Hence, by standardizing the Xs only, you can see the relative importance of the Xs. We see that a 1 standard deviation increase in gpa produces, on average, a 1.319 increase in the log odds of getting an A. (To get the X-Standardized coefficient, just multiple b_k by the standard deviation of x_k , e.g. for gpa $2.82611 * .4667 = 1.319$.)

Y-Standardization. You can also standardize Y* only. The `listcoef` column labeled `bStdY` gives you the coefficients from when Y* is standardized but X is not. A 1 unit increase in gpa produces, on average, a 1.0525 standard deviation increase in Y*. To get the Y-standardized coefficient, just divide b_k by the standard deviation of Y*, e.g. for gpa $2.82611/2.685 = 1.0525$.

If you don't include the `std` parameter, after a logistic regression `listcoef` does a variation of X-standardization, showing you the odds ratios (i.e. the factor change in the odds as X increases):

```
. listcoef
logit (N=32): Factor Change in Odds
Odds of: 1 vs 0
-----
      grade |      b      z  P>|z|  e^b  e^bStdX  SDofX
-----+-----
      gpa |  2.82611  2.238  0.025  16.8797  3.7396  0.4667
      tuce |  0.09516  0.672  0.501   1.0998  1.4496  3.9015
      psi |  2.37869  2.234  0.025  10.7907  3.3165  0.5040
-----
```

This tells you that a 1 unit increase in gpa multiplies the odds of success by 16.8797. A 1 standard deviation increase in gpa multiplies the odds by 3.7396. (Recall that the X-standardized coefficient is 1.3190; $\exp(1.3190) = 3.7396$.)

Discussion

The question remains, why would you want to do any of the above? The usual argument for using standardized coefficients is that they provide a means for comparing the effects of variables which may be measured in a wide variety of metrics. This is true here as well. So, for example, you can see that a 1 standard deviation increase in gpa produces more change in the log odds of getting an A than does a 1 standard deviation increase in tuce.

There are, however, some unique concerns when using logistic regression and other GLMs. Unlike Y in OLS regression, the variance of Y* is not fixed; it will change as you add more variables to the model. Further, as you add variables, coefficients will change even if the new variables are uncorrelated with the old ones. This makes comparisons of coefficients across models problematic. Some authors (e.g. Winship & Mare, ASR 1984) therefore recommend Y-Standardization or Full-Standardization.

An example will illustrate this. I have constructed a data set such that x1 and x2 are uncorrelated with each other. Both have strong effects on y. ybinary is a dichotomized version of y, where y values above 0 are recoded to 1 and values of 0 and below are recoded to 0. Let's compare the results of OLS regression and logistic regression.

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/standardized.dta
```

```
. corr, means
```

```
(obs=500)
```

Variable	Mean	Std. Dev.	Min	Max
y	5.51e-07	3.000001	-8.508021	7.981196
ybinary	.488	.5003566	0	1
x1	-2.19e-08	2	-6.32646	6.401608
x2	3.57e-08	3	-10.56658	9.646875

	y	ybinary	x1	x2
y	1.0000			
ybinary	0.7923	1.0000		
x1	0.6667	0.5248	1.0000	
x2	0.6667	0.5225	0.0000	1.0000

```
. qui reg y x1
```

```
. listcoef, std
```

```
regress (N=500): Unstandardized and Standardized Estimates
```

```
Observed SD: 3.0000014
```

```
SD of Error: 2.2383128
```

	y	b	t	P> t	bStdX	bStdY	bStdXY	SDofX
x1		1.00000	19.960	0.000	2.0000	0.3333	0.6667	2.0000

```
. qui reg y x2
```

```
. listcoef, std
```

```
regress (N=500): Unstandardized and Standardized Estimates
```

```
Observed SD: 3.0000014
```

```
SD of Error: 2.2383131
```

```
-----+-----  
      y |      b      t    P>|t|    bStdX    bStdY    bStdXY      SDofX  
-----+-----  
      x2 |  0.66667  19.960  0.000    2.0000    0.2222    0.6667      3.0000  
-----+-----
```

```
. qui reg y x1 x2
```

```
. listcoef, std
```

```
regress (N=500): Unstandardized and Standardized Estimates
```

```
Observed SD: 3.0000014
```

```
SD of Error: 1.0020108
```

```
-----+-----  
      y |      b      t    P>|t|    bStdX    bStdY    bStdXY      SDofX  
-----+-----  
      x1 |  1.00000  44.587  0.000    2.0000    0.3333    0.6667      2.0000  
      x2 |  0.66667  44.587  0.000    2.0000    0.2222    0.6667      3.0000  
-----+-----
```

As we see, in an OLS regression, when x_1 and x_2 are uncorrelated with each other, their metric and standardized effects are the same in the bivariate regressions as they are when y is regressed on both x 's simultaneously. (Basically, this is the special case of omitted variable bias: when the x 's are uncorrelated with each other, leaving one x out does not bias the estimated effect of the other.)

Compare this now to the results of a logistic regression:

```
. quietly logit ybinary x1
```

```
. listcoef, std
```

```
logit (N=500): Unstandardized and Standardized Estimates
```

```
Observed SD: .50035659
```

```
Latent SD: 2.3395663
```

```
Odds of: 1 vs 0
```

```
-----+-----  
ybinary |      b      z    P>|z|    bStdX    bStdY    bStdXY      SDofX  
-----+-----  
      x1 |  0.73887  10.127  0.000    1.4777    0.3158    0.6316      2.0000  
-----+-----
```

```
. quietly logit ybinary x2
```

```
. listcoef, std
```

```
logit (N=500): Unstandardized and Standardized Estimates
```

```
Observed SD: .50035659
```

```
Latent SD: 2.3321875
```

```
Odds of: 1 vs 0
```

ybinary	b	z	P> z	bStdX	bStdY	bStdXY	SDofX
x2	0.48868	10.134	0.000	1.4660	0.2095	0.6286	3.0000

```
. quietly logit ybinary x1 x2
```

```
. listcoef, std
```

```
logit (N=500): Unstandardized and Standardized Estimates
```

```
Observed SD: .50035659
```

```
Latent SD: 5.3368197
```

```
Odds of: 1 vs 0
```

ybinary	b	z	P> z	bStdX	bStdY	bStdXY	SDofX
x1	1.78923	9.815	0.000	3.5785	0.3353	0.6705	2.0000
x2	1.17314	9.714	0.000	3.5194	0.2198	0.6595	3.0000

In the bivariate logistic regressions, the unstandardized coefficients for x1 and x2 are about .74 and .49 respectively; but when x1 and x2 are both in the equation, the coefficients are dramatically different, 1.79 and 1.17!

If we saw those kinds of changes in an OLS regression, we'd probably start thinking that suppressor effects were present, e.g. something like this might occur if x1 and x2 were negatively correlated while both had positive effects on y. But, since, in this example, x1 and x2 are uncorrelated, that is obviously not what is going on here. Rather, note how the standard deviation of Y* fluctuates from one logistic regression to the next; it is about 2.34 in each of the bivariate logistic regressions and 5.34 in the multivariate logistic regression. It is because the variance of Y* changes that the coefficients change so much when you go from one regression to the next. (By way of analogy, if in one OLS regression income was measured in dollars, and in another it was measured in thousands of dollars, the coefficients would be very different.)

Compare this to the changes in the Y-Standardized and Fully-Standardized coefficients. The Y-Standardized coefficients for x1 and x2 are .3158 and .2095 in the bivariate regressions; in the multivariate regressions they are .3353 and .2198. Changes in the standardized coefficients are far less than the changes in the non-standardized coefficients. If your goal is to compare coefficients across nested models, it is probably better to use Y-Standardized or Fully-Standardized coefficients.

Another way of thinking about this: I've been highly critical of the use of standardized coefficients in OLS regression. But, in logistic regression, you are basically doing a different type of standardization: the variance of the residuals is fixed at $\pi^2/3$, or about 3.29. (Since you don't actually observe Y^* , you have to identify its variance in some way.) Hence, logistic regression and other GLMs already have some of the problems of standardized coefficients inherently built into them. Given that an arbitrary type of standardization is already going on anyway, Y-Standardization or Full-Standardization may be superior for your purposes.

To elaborate further: Often researchers present a hierarchy of models, e.g. they will estimate a model with x_1 - x_3 included, then in a second model they will add x_4 - x_6 , then the third will add x_7 - x_9 , etc. As part of the discussion of the results, it might be noted how the effects of early variables decline or increase as additional variables are added, e.g. "The effect of race declines once income is controlled for." Such comparisons of coefficients are potentially misleading in a logistic regression; coefficient estimates can change, not just because the effect of a variable increases or decreases as others are controlled, but because $V(Y^*)$ is changing as new variables are added. Ergo, some of the things we are used to doing with metric coefficients in OLS regression are not legit when doing logistic regression. If we are interested in discussing how the effects of variables change as controls are added (and we may or may not be), Y-Standardized coefficients are probably better.

Finally, there is nothing that says you couldn't use variations of the above methods. For example, with the X variables, you could use combinations of standardized and unstandardized variables. Many people don't like to standardize dummy variables, which only have values of 0 and 1, because a "one standard deviation increase" isn't something that could actually happen with such a variable. Ergo, you might want to leave the dummy variables unstandardized while standardizing continuous X variables.

Conclusion: in OLS regression, I am not very fond of standardized coefficients – I think they can often take something that is fairly intuitive and easy to understand (e.g. the effect of education on income) and make it much less intuitive, and there are other problems as well. If comparing the effects of the X's is a goal, X-Standardization may be as good or better as Full-Standardization.

But, with logistic regression and other GLMs, some of the problems of standardized coefficients are inherently built in to the model, because these also do a type of standardization: the residuals are standardized to have a variance of 3.29. This has many advantages, but it also has drawbacks you need to be aware of. Depending on your purposes, you may find that a different type of standardization, e.g. Y-Standardization, is better. (Of course, then you have to explain to the reader what those are!) In any event, you should be aware of the potential pitfalls, e.g. your discussion should be careful about making comparisons of coefficients across models that may not be valid.

Additional Comments

RWLS as an alternative to OLS (or why we should be grateful OLS isn't more like logistic regression). In logistic regression, the variance of ε_{y^*} is fixed at 3.29. In OLS, the Total Sums of Squares stays the same as you add variables, but the regression and error sums of squares can vary. Suppose (perhaps out of some insane desire for consistency) we had a variation of OLS which did something similar to logistic regression, e.g. it standardized Y so that the mean square error (MSE) in a regression was always 3.29. We'll call this alternate-universe version of ordinary least squares RWLS (Rich Williams Least Squares). How would regression coefficients behave under such an alternate method? The following examples show what would happen.

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/standardized.dta
. quietly reg y x1
. gen double ystar = y / sqrt((e(rss)/e(df_r))) * sqrt(3.29)
. reg ystar x1, beta
```

Source	SS	df	MS	Number of obs =	500

Model	1310.73621	1	1310.73621	F(1, 498)	= 398.40
Residual	1638.42	498	3.29	Prob > F	= 0.0000

Total	2949.15621	499	5.91013269	R-squared	= 0.4444

				Adj R-squared	= 0.4433
				Root MSE	= 1.8138

ystar	Coef.	Std. Err.	t	P> t	Beta

x1	.8103589	.0405992	19.96	0.000	.6666667
_cons	4.65e-07	.0811172	0.00	1.000	.

```
. quietly reg y x2
. replace ystar = y / sqrt((e(rss)/e(df_r))) * sqrt(3.29)
```

(500 real changes made)

```
. reg ystar x2, beta
```

Source	SS	df	MS	Number of obs =	500

Model	1310.73549	1	1310.73549	F(1, 498)	= 398.40
Residual	1638.42	498	3.29	Prob > F	= 0.0000

Total	2949.15549	499	5.91013125	R-squared	= 0.4444

				Adj R-squared	= 0.4433
				Root MSE	= 1.8138

ystar	Coef.	Std. Err.	t	P> t	Beta

x2	.5402391	.0270661	19.96	0.000	.6666666
_cons	4.28e-07	.0811172	0.00	1.000	.

```
. quietly reg y x1 x2
. replace ystar = y / sqrt((e(rss)/e(df_r))) * sqrt(3.29)
```

(500 real changes made)

```
. reg ystar x1 x2, beta
```

Source	SS	df	MS		
Model	13081.0321	2	6540.51606	Number of obs =	500
Residual	1635.13	497	3.29	F(2, 497) =	1988.00
Total	14716.1621	499	29.4913069	Prob > F =	0.0000
				R-squared =	0.8889
				Adj R-squared =	0.8884
				Root MSE =	1.8138

ystar	Coef.	Std. Err.	t	P> t	Beta
x1	1.810197	.0405992	44.59	0.000	.6666667
x2	1.206798	.0270661	44.59	0.000	.6666666
_cons	9.95e-07	.0811172	0.00	1.000	.

In the RWLS bivariate regressions, the unstandardized coefficients for x1 and x2 are about .81 and .54 respectively; but when both x1 and x2 are both in the equation, the coefficients are dramatically different, 1.81 and 1.21 (unlike OLS, where they didn't change at all). Further, the variance of ystar (as shown by the MS Total) is about 5.91 in each of the bivariate regressions but zooms to 29.49 in the multivariate regression. However, the standardized coefficients are the same throughout. In short, *if OLS was more like logistic regression, where the error variance was fixed instead of free to vary, we'd see the same sort of oddities in the parameter estimates as we went from one model to the next as we did with logistic regression.*

I've often noted the evils of standardized coefficients in OLS. But, if I had to choose between RWLS and standardized coefficients, I might grudgingly go with standardized coefficients – at least they don't create the same kind of moving target with $V(Y)$ that RWLS does. Luckily, we don't have to use RWLS! (But unfortunately, I'll have to keep looking for a statistical technique I can name after myself – this one is doomed to either forever live in infamy or else be quickly forgotten.)

Question: In logistic regression, why does the variance of y^* increase as you add more variables?

There are at least two ways of answering this. First, it increases because it has to. In OLS regression, SST (Total Sums of Squares) stays the same as you add more variables; but the Regression Sums of Squares is free to increase while the Error Sums of Squares makes a corresponding decrease. But, in logistic regression, $V(\epsilon_{y^*})$ is fixed at 3.29. Since the error variance can't go down, the explained variance (and hence the total variance) has to go up as you add more variables.

Second, the variance of y^* changes because your estimation of the probability of success gets better and better as you add more variables, and better estimation in turn leads to more variability in the estimates. Suppose you had a single dichotomous IV, e.g. gender. All men would have the same probability of success (and hence the same logit) and all women would also share the same probability/logit. For example, men might have a 40% probability of success while women had a 50% chance.

As you add more variables, however, there can start to be more variability in the logits. So, among the men, after adding another variable the probabilities might now range between 35% and 45%, while for women the range might be 45% to 55%. Hence, instead of the probabilities only ranging from 40% to 50%, adding another variable could cause the probabilities to range from 35% to 55%, and the logits would vary more as well. If you are doing really really well, adding more variables might get you to the point where the probabilities ranged from .00001 to .9999, and the logits ranged from something like -10 to 10.

To put it another way: When relevant variables are missing, you will overestimate the probability of success for some (e.g. some men will have less than a 40% chance of success) while underestimating it for others (e.g. some women will have better than a 60% chance). As your model improves you will get more accurate estimates of the probability of success which in turn will result in greater variability in y^* .

Question: How serious is the problem in practice? How can you avoid the problem?

- It won't be an issue at all if researchers don't attempt to compare coefficients across nested models. Indeed, it is very common to report model fit statistics (e.g. model chi-square, BIC, chi-square contrasts) for intermediate models, and to give coefficient estimates only for the final model. If you aren't going to focus much on changes in coefficients anyway, this is a very good strategy, and would probably be my first choice in most cases. Presenting a lot of unnecessary and potentially misleading coefficient estimates may do you more harm than good.
- If you do want to compare coefficients, it might not be that big of an issue if
 - $V(y^*)$ doesn't change that much from one model to the next.
 - Coefficients decline as you add variables rather than increase. In this case, you are actually underestimating the amount of decline. Conversely, if coefficients increase as you add more variables, you have to be careful that any argument you want to make about suppressor effects is valid.
- Nonetheless, if you are going to compare coefficients, it seems best to use y-standardization, or at least check to make sure y-standardization would not change your conclusions.

My guess is that the problem doesn't come up that much in practice, partly because researchers using these methods seem less inclined to present detailed results for each model. Also, any comparisons that are made tend to focus on changes in statistical significance rather than changes in magnitudes of effects. But, if you see articles where researchers appear to be making invalid comparisons across models, I would be interested in hearing about them.

Question: How well known is this problem? Are there any citations on this?

As noted before, Winship and Mare alluded to the need for y-standardization back in 1984, and they in turn were citing earlier work. Nonetheless, I don't think the problem is widely known and understood. If you want to compare coefficients across nested models and justify your use of y-standardization (and you want to quote something besides just a handout you found on the Internet) Mare briefly discusses the issue:

When the error variance is fixed, it is also inappropriate to make within sample comparisons among the coefficients for a given covariate across equations with varying subsets of covariates. In this case, the total variance of the latent dependent variable and thus the scale of the estimated coefficients vary from model to model as a function of the different regressors that are included. Fixing the variance of the latent dependent variable avoids this problem. It does not, however, avoid the problems of comparison across samples and across dependent variables.

Source: *Response: Statistical Models of Educational Stratification—Hauser and Andrew's Models for School Transitions*. Robert D. Mare. [Sociological Methodology 2006](#).

Question: Suppose you had never heard of, or didn't believe in, the latent variable model. How would you explain these results using the nonlinear probability model?

Good question. I don't know. I've seen other authors provide these sorts of dual proofs for other topics, but I am not sure how to do it in this case. Any ideas are welcome.