

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 i.psi, nolog
```

```
Logistic regression                               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
1.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(27):	25.779	LR(3):	15.404
		Prob > LR:	0.002
McFadden's R2:	0.374	McFadden's Adj R2:	0.131
ML (Cox-Snell) R2:	0.382	Cragg-Uhler(Nagelkerke) R2:	0.528
McKelvey & 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:	1.118	AIC*n:	35.779
BIC:	-67.796	BIC':	-5.007
BIC used by Stata:	39.642	AIC used by Stata:	33.779

```
. 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
1.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 later). 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 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 multiply 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; $\exp(1.3190) = 3.7396$.) See the help for `listcoef` for other options that may be useful.

Discussion. The usual argument for using standardized coefficients is that they provide a means for comparing the effects of variables measured in different metrics. This is true here as well. So, for example, you can see that a 1 standard deviation (SD) increase in `gpa` produces more change in the log odds of getting an A than does a 1 SD increase in `tuce`. Nevertheless, standardized effects tend to be looked down upon. It makes no sense to think about a one SD increase in a dummy variable like `gender`. Even for continuous variables, standardized coefficients are not very intuitive, e.g. how many of us think in terms of standard deviations? Worse, they can be very misleading. For example, if the standard deviations of variables differ across groups, the standardization of variables will also differ, causing coefficients to not be comparable across groups (e.g. in one group X might get divided by 10 while in another it gets divided by 7.)

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. We discuss this further in a later handout.

Appendix: Standardized Coefficients in OLS Regression

If you run `listcoef` after the `regress` command, the fully standardized coefficients are the same as the regression standardized coefficients, e.g.

```
. webuse nhanes2f, clear
. reg weight height age female black, beta
```

Source	SS	df	MS	Number of obs =	10337
Model	620082.606	4	155020.652	F(4, 10332) =	881.52
Residual	1816944.64	10332	175.856044	Prob > F =	0.0000
				R-squared =	0.2544
				Adj R-squared =	0.2542
Total	2437027.25	10336	235.7805	Root MSE =	13.261

weight	Coef.	Std. Err.	t	P> t	Beta
height	.7485279	.01966	38.07	0.000	.4709032
age	.1237255	.0078948	15.67	0.000	.1387257
female	-1.540187	.3721392	-4.14	0.000	-.0500913
black	3.679295	.4256284	8.64	0.000	.0734762
_cons	-59.05337	3.563342	-16.57	0.000	.

```
. listcoef, std help
```

regress (N=10337): Unstandardized and Standardized Estimates

```
Observed SD: 15.355146
SD of Error: 13.261072
```

weight	b	t	P> t	bStdX	bStdY	bStdXY	SDofX
height	0.74853	38.074	0.000	7.2308	0.0487	0.4709	9.6600
age	0.12373	15.672	0.000	2.1302	0.0081	0.1387	17.2168
female	-1.54019	-4.139	0.000	-0.7692	-0.1003	-0.0501	0.4994
black	3.67929	8.644	0.000	1.1282	0.2396	0.0735	0.3066

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

Note that, in OLS, while full standardization is frequently done, X-Standardization alone is enough to achieve the goal of comparing the effects of Xs measured in different metrics, and may be easier to interpret since Y is left in its original metric. So, for example, we can see that a 1 standard deviation in height results, in average, on a 7.23 kilogram increase in weight, whereas a 1 standard deviation increase in age results in an average increase of 2.13 kilograms.