Standard errors for regression coefficients; Multicollinearity

Standard errors. Recall that b_k is a *point* estimate of β_k. Because of sampling variability, this estimate may be too high or too low. s_{bk}, the standard error of b_k, gives us an indication of how much the point estimate is likely to vary from the corresponding population parameter. We will now broaden our earlier discussion.

Let H = the set of all the X (independent) variables.

Let G_k = the set of all the X variables *except* X_k.

The following formulas then hold:

General case:	$s_{b_k} = \frac{s_e}{\sqrt{(1-R^2_{X_kG_k}) * s^2_{X_k} * (N-1)}}$ $= \sqrt{\frac{1-R^2_{YH}}{(1-R^2_{X_kG_k}) * (N-K-1)}} * \frac{s_y}{s_{X_k}}$	The first formula uses the standard error of the estimate. The second formula makes it clearer how standard errors are related to R^2.
2 IV case	$s_{b_k} = \frac{s_e}{\sqrt{(1-R^2_{12}) * s^2_{X_k} * (N-1)}}$ $= \sqrt{\frac{1-R^2_{Y12}}{(1-R^2_{12}) * (N-K-1)}} * \frac{s_y}{s_{X_k}}$	When there are only 2 IVs, $R^2_{XkGk} = R^2_{12}$.
1 IV case	$s_b = \frac{s_e}{\sqrt{s^2_X * (N-1)}}$ $= \sqrt{\frac{1-R^2}{(N-K-1)}} * \frac{s_Y}{s_X}$	When there is only 1 IV, $R^2_{XkGk} = 0$.

For example, if K = 5, then R_{YH}^2 is the multiple R^2 obtained by regression Y on X_1, X_2, X_3, X_4, and X_5; and, if we wanted to know s_{b3} (i.e. the standard error for X_3) then R_{X3G3}^2 would be the multiple R^2 obtained by regressing X_3 on X_1, X_2, X_4, and X_5. Note that, when there are 2 independent variables, $R_{X1G1}^2 = R_{X2G2}^2 = R_{12}^2$.

Question. Suppose K = 1, i.e. there is only 1 independent variable. What is the correct formula then?

Answer. When K = 1, $R_{XkGk}^2 = 0$ (because there are no other X's to regress on X_1). The general formula then becomes the same as the formula we presented when discussing bivariate regression.

Question. What happens as R_{XkGk}^2 gets bigger and bigger?

Answer. As R_{XkGk}^2 gets bigger and bigger, the denominator in the above equations gets smaller and smaller, hence the standard errors get larger and larger. For example:

If $R_{12}^2 = 0$, and nothing else changes, then,

$$s_{b_1} = \frac{s_e}{\sqrt{(1-R_{12}^2)*s_{X_1}^2*(N-1)}} = \frac{4.08}{\sqrt{(1-0)*20.05*(19)}} = \frac{4.08}{\sqrt{380.95}} = .209$$

$$= \sqrt{\frac{1-R_{Y12}^2}{(1-R_{12}^2)*(N-K-1)}} * \frac{s_y}{s_{X_1}} = \sqrt{\frac{1-.84498}{(1-0)*(17)}} * \frac{9.788}{4.478} = .209$$

$$s_{b_2} = \frac{s_e}{\sqrt{(1-R_{12}^2)*s_{X_2}^2*(N-1)}} = \frac{4.08}{\sqrt{(1-0)*29.82*(19)}} = \frac{4.08}{\sqrt{566.55}} = .171$$

$$= \sqrt{\frac{1-R_{Y12}^2}{(1-R_{12}^2)*(N-K-1)}} * \frac{s_y}{s_{X_2}} = \sqrt{\frac{1-.84498}{(1-0)*(17)}} * \frac{9.788}{5.461} = .171$$

If $R_{12}^2 = .25$ and nothing else changes, then,

$$s_{b_1} = \frac{s_e}{\sqrt{(1-R_{12}^2)*s_{X_1}^2*(N-1)}} = \frac{4.08}{\sqrt{(1-.25)*20.05*(19)}} = \frac{4.08}{\sqrt{285.71}} = .241$$

$$= \sqrt{\frac{1-R_{Y12}^2}{(1-R_{12}^2)*(N-K-1)}} * \frac{s_y}{s_{X_1}} = \sqrt{\frac{1-.84498}{(1-.25)*(17)}} * \frac{9.788}{4.478} = .241$$

$$s_{b_2} = \frac{s_e}{\sqrt{(1-R_{12}^2)*s_{X_2}^2*(N-1)}} = \frac{4.08}{\sqrt{(1-.25)*29.82*(19)}} = \frac{4.08}{\sqrt{424.935}} = .198$$

$$= \sqrt{\frac{1-R_{Y12}^2}{(1-R_{12}^2)*(N-K-1)}} * \frac{s_y}{s_{X_2}} = \sqrt{\frac{1-.84498}{(1-.25)*(17)}} * \frac{9.788}{5.461} = .198$$

Similarly, you can show that, if $R_{12}^2 = .5$, then $s_{b1} = .295$ and $s_{b2} = .242$.

Question. Suppose R_{XkGk}^2 is very large, say, close to 1. What happens then?

Answer. If $R_{XkGk}^2 = 1$, then $1 - R_{XkGk}^2 = 0$, which means that the standard error becomes infinitely large. Ergo, the closer R_{XkGk}^2 is to 1, the bigger the standard error gets. Put another way, the more correlated the X variables are with each other, the bigger the standard errors

become, and the less likely it is that a coefficient will be statistically significant. This is known as the problem of **multicollinearity**.

Intuitively, the reason this problem occurs is as follows: The more highly correlated independent variables are, the more difficult it is to determine how much variation in Y each X is responsible for. For example, if X_1 and X_2 are highly correlated (which means they are very similar to each other) it is difficult to determine whether X_1 is responsible for variation in Y, or whether X_2 is. As a result, the standard errors for both variables become very large. In our current example, if $R_{12}^2 = .95$, then $s_{b1} = .933$ and $s_{b2} = .765$. Note that, under these conditions, neither coefficient would be significant at the .05 level, even though their combined effects are statistically significant.

Comments:

1. It is possible for all independent variables to have relatively small mutual correlations and yet to have some multicollinearity among three or more of them. The multiple correlation R_{XkGk} can indicate this.

2. When multicollinearity occurs, the least-squares estimates are still unbiased and efficient. The problem is that the estimated standard errors of the coefficients tend to be inflated. That is, the standard error tends to be larger than it would be in the absence of multicollinearity because the estimates are very sensitive to changes in the sample observations or in the model specification. In other words, including or excluding a particular variable or certain observations may greatly change the estimated coefficients.

3. One way multicollinearity can occur is if you accidentally include the same variable twice, e.g. height in inches and height in feet. Another common error occurs when one of the X's is computed from the other X's (e.g. Family income = Wife's income + Husband's income), and the computed variable and the variables used to compute it are all included in the regression equation. Improper use of dummy variables (which we will discuss later) can also lead to perfect collinearity. These errors are all avoidable. However, other times, it just happens to be the case that the X variables are naturally highly correlated with each other.

4. Many computer programs for multiple regression help guard against multicollinearity by reporting a "tolerance" figure for each of the variables entering into a regression equation. This tolerance is simply the proportion of the variance for the variable in question that is not due to other X variables; that is, **Tolerance $X_k = 1 - R_{XkGk}^2$**. A tolerance value close to 1 means you are very safe, whereas a value close to 0 shows that you run the risk of multicollinearity, and possibly no solution, by including this variable.

Note, incidentally, that the tolerance appears in the denominator of the formulas for the standard errors. As the tolerance gets smaller and smaller (i.e. as multicollinearity increases) standard errors get bigger and bigger.

Also useful is the **Variance Inflation Factor (VIF)**, which is the reciprocal of the tolerance. This shows us how much the variances are inflated by multicollinearity, e.g. if the VIF is 1.44, multicollinearity is causing the variance of the estimate to be 1.44 times larger than it would be if the independent variables were uncorrelated (meaning that the standard error is 1.44 = 1.2 times larger).

5. There is no simple means for dealing with multicollinearity (other than to avoid the sorts of common mistakes mentioned above.) Some possibilities:

a. Exclude one of the X variables - although this might lead to specification error

b. Find another indicator of the concept you are interested in, which is not collinear with the other X's.

c. Put constraints on the effects of variables, e.g. require that 2 or more variables have equal effects (or effects of equal magnitude but opposite direction.) For example, if years of education and years of job experience were highly correlated, you might compute a new variable which was equal to EDUC + JOBEXP, and use it instead.

d. Collect a larger sample, since larger sample sizes reduce the problem of multicollinearity by reducing standard errors.

e. In general, be aware of the possible occurrence of multicollinearity, and know how it might distort your parameter estimates and significance tests.

Here again is an expanded printout from SPSS that shows the tolerances and VIFs:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	Std. Error			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-7.097	3.626			-1.957	.067	-14.748	.554					
	EDUC	1.933	.210	.884	.096	9.209	.000	1.490	2.376	.846	.913	.879	.989	1.012
	JOBEXP	.649	.172	.362	.096	3.772	.002	.286	1.013	.268	.675	.360	.989	1.012

a. Dependent Variable: INCOME

Another example. Let's take another look at one of your homework problems. We will examine the tolerances and show how they are related to the standard errors.

```
            Mean  Std Dev  Variance  Label

XHWORK    3.968    2.913     8.484  TIME ON HOMEWORK PER WEEK
XBBSESRW  -.071     .686      .470  SES COMPOSITE SCALE SCORE
ZHWORK    3.975    2.930     8.588  TIME ON HOMEWORK PER WEEK
XTRKACAD   .321     .467      .218  X IN ACADEMIC TRACK

N of Cases =  9303

Equation Number 1    Dependent Variable..   XHWORK   TIME ON HOMEWORK PER WEEK

Multiple R           .40928
R Square             .16751
Standard Error      2.65806

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           3         13219.80246       4406.60082
Residual          9299         65699.89382          7.06526

F =     623.69935       Signif F =  .0000

---------------------------------- Variables in the Equation -----------------------------------

Variable              B        SE B     95% Confdnce Intrvl B       Beta    SE Beta   Correl Part Cor

XBBSESRW        .320998     .042126      .238422      .403575    .075555    .009915  .179292  .072098
ZHWORK          .263356     .009690      .244363      .282350    .264956    .009748  .325969  .257166
XTRKACAD       1.390122     .062694     1.267227     1.513017    .222876    .010052  .303288  .209795
(Constant)     2.496854     .049167     2.400475     2.593233

---------- Variables in the Equation ----------
Variable    Partial  Tolerance         VIF         T  Sig T

XBBSESRW    .078774    .910596       1.098     7.620  .0000
ZHWORK      .271284    .942060       1.062    27.180  .0000
XTRKACAD    .224088    .886058       1.129    22.173  .0000
(Constant)                                    50.783  .0000
```

To simplify the notation, let Y = XHWORK, X_1 = XBBSESRW, X_2 = ZHWORK, X_3 = XTRKACAD. The printout tells us

N = 9303, SSE = 65699.89382, s_e = 2.65806, R_{Y123}^2 = .16751,
s_y = 2.913, s_1 = .686, s_2 = 2.930, s_3 = .467,

Tolerance X_1 = .910596 ==> R_{X1G1}^2 = 1 - .910596 = .089404
Tolerance X_2 = .942060 ==> R_{X2G2}^2 = 1 - .942060 = .057940
Tolerance X_3 = .886058 ==> R_{X3G3}^2 = 1 - .886058 = .113942

The high tolerances and the big sample size strongly suggest that we need not be worried about multicollinearity in this problem.

We will now compute the standard errors, using the information about the tolerances.

$$s_{b_1} = \sqrt{\frac{1 - R^2_{YH}}{(1 - R^2_{X_1G_1}) * (N - K - 1)}} * \frac{s_y}{s_{x_1}} =$$

$$= \sqrt{\frac{1 - .168}{.91 * 9299}} * \frac{2.913}{.686} = .009916 * \frac{2.913}{.686} = .0421$$

$$s_{b_2} = \sqrt{\frac{1 - R^2_{YH}}{(1 - R^2_{X_2G_2}) * (N - K - 1)}} * \frac{s_y}{s_{x_2}} =$$

$$= \sqrt{\frac{1 - .168}{.94 * 9299}} * \frac{2.913}{2.93} = .009756 * \frac{2.913}{2.930} = .0097$$

$$s_{b_3} = \sqrt{\frac{1 - R^2_{YH}}{(1 - R^2_{X_3G_3}) * (N - K - 1)}} * \frac{s_y}{s_{x_3}} = s_{b_{3'}} * \frac{s_y}{s_{x_3}}$$

$$= \sqrt{\frac{1 - .168}{.886 * 9299}} * \frac{2.913}{.467} = .010049 * \frac{2.913}{.467} = .0627$$

As you can see, our computed standard errors are consistent with SPSS.

We will have a more extensive discussion of multicollinearity in Stats II.