Comments on standardizing path diagrams: what are the parameters?

Let
\[ Y_i = a + bU_i + cV_i + \delta_i \]  
and
\[ Z_i = \alpha + \beta Y_i + \epsilon_i. \]

Take \( U_i, V_i \) as data, with mean 0, variance 1, and correlation \( r \). The \( \delta_i \) are IID with mean 0 and variance \( \sigma^2 \). The \( \epsilon_i \) are IID with mean 0 and variance \( \tau^2 \), independent of the \( \delta_i \). (See exercise 5C6 in *Statistical Models*.) Let \( s_Y \) be the standard deviation of \( \{Y_1, \ldots, Y_n\} \). If we standardize the \( Y_i \), then (i) we’re dividing by a random variable, \( s_Y \); and (ii), the \( \delta_i \) get dependent. So, what are the parameters?

One solution is to standardize \( Y \) at the population level. First,
\[
E \left[ \frac{1}{n} \sum_{i=1}^{n} (Y_i - a)^2 \right] = b^2 + c^2 + 2br + \sigma^2 = \theta^2,
\]
say. So, replace \( Y_i \) by \( \eta_i = (Y_i - a)/\theta \). We have
\[
\eta_i = \frac{b}{\theta} U_i + \frac{c}{\theta} V_i + \frac{\delta_i}{\theta} \tag{3}
\]
Thus \( E(\eta) = 0 \) and \( E(\eta^2) = 1 \), although
\[
E(\eta_i) = \frac{bU_i + cV_i}{\theta} \neq 0 \quad \text{and} \quad E(\eta^2_i) = \frac{(bU_i + cV_i)^2 + \sigma^2}{\theta^2} \neq 1.
\]

Standardization is “on the average,” over the whole population. Note that (3) is a bona fide regression equation, with all the usual assumptions on the errors. Fitting the standardized equation can be viewed as estimating \( b/\theta, c/\theta, \sigma^2/\theta^2 \). The estimates will suffer from ratio estimator bias, due to division by the random \( s_Y \).

The trick for (2) is the same. First, replace \( Z_i \) by
\[
Z^*_i = (Z_i - \alpha - a\beta)/\theta.
\]
We get the regression equation
\[
Z^*_i = \beta \eta_i + \frac{\epsilon_i}{\theta}
\]
Let
\[
\phi^2 = E \left[ \frac{1}{n} \sum_{i=1}^{n} Z^*_i^2 \right] = \beta^2 + \frac{\tau^2}{\theta^2}.
\]
Finally, replace $Z^*_i$ by $\zeta_i = Z^*_i / \phi$. When standardized at the population level, (2) becomes

$$\zeta_i = \frac{\beta}{\phi} \eta_i + \frac{\epsilon_i}{\theta \phi}$$

(4)

Again, $E(\zeta_i) \neq 0$ and $E(\zeta_i^2) \neq 1$, so the standardization only applies “on average:” $E(\zeta) = 0$ and $E(\zeta^2) = 1$. But (4) is a legitimate regression equation.

In the leading special case, $U_i, V_i, \delta_i, \epsilon_i$ are IID in $i$. We can center $U_i, V_i$ at their expected values and divide by the respective standard deviations. The endogenous variables $Y_i, Z_i$ now have expectation 0. Division by the respective SEs achieves standardization—at the population level—for each $i$. The sample will not be standardized exactly, due to random error. Again, standardizing the sample leads to a minor ratio-estimation bias, with a minor gain on the variance side since intercepts do not need to be estimated.

Also see exercise 5C6 on pp. 84–85 of Statistical Models.