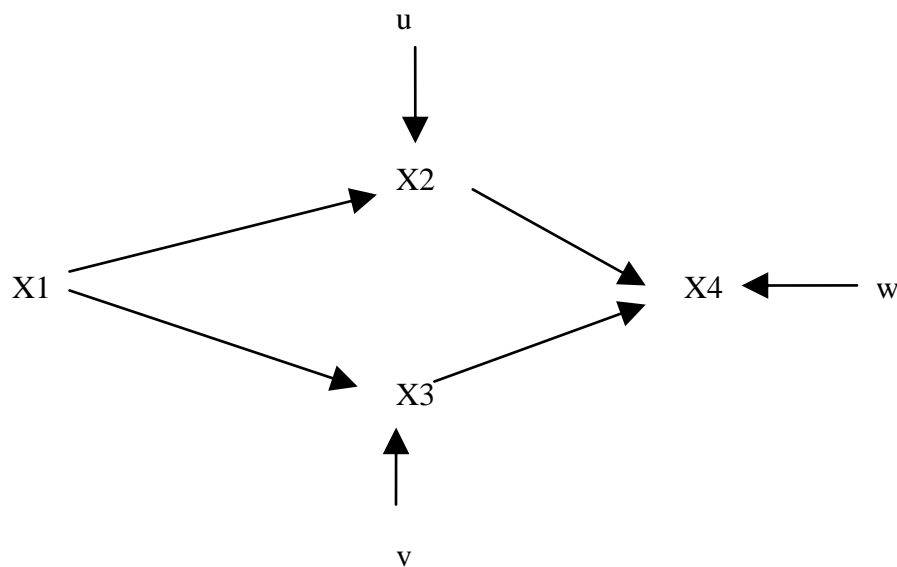


Computing R²/Evils of R²

COMPUTING R². Here are some of the many formulas for R²: Our knowledge of path analysis now makes it possible to prove many of these formulas. See the optional appendix if you are interested. NOTE: I use the notation b'_k for the standardized coefficients and b_k for the non-standardized, aka metric coefficients. p_k is another notation often used for the standardized (aka path) coefficients.

$R^2 = SSR/SST$	Explained sum of squares over total sum of squares, i.e. the ratio of the explained variability to the total variability.
$R^2 = \frac{F * K}{(N - K - 1) + (F * K)}$	This can be useful if F, N, and K are known
$R^2 = \sum_{k=1}^K b'_k r_{yk}$ <p style="text-align: center;">Also,</p> $R^2 = \sum_{i=1}^K \sum_{j=1}^K b'_i b'_j r_{ij}$	These formulas uses the standardized coefficients.and the zero-order correlations between y and the x's. These (esoteric) formulas can be useful when doing path analysis.
<p>Two IV case only:</p> $R^2 = b_1'^2 + b_2'^2 + 2b_1' b_2' r_{12}$	This is a special case of the last formula.
<p>One IV case only:</p> $R^2 = b'^2$	Remember that, in standardized form, correlations and covariances are the same.
<p>One IV case only:</p> $R^2 = \frac{V(\hat{Y})}{V(Y)} = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)}$	We'll use this formula in the rest of this handout. Technically I should use rho ² (ρ ²) here or else put hats over the other parameters (to distinguish between population parameters and sample estimates) but since that seems to confuse some people I won't!

EVILS OF R^2 (In General) While R^2 has its uses, it is far more important to have a correctly specified model than it is to have a large R^2 . For example, consider the following model:



- You can increase R^2 by regressing a variable on variables that are causally subsequent to it, e.g. regress X_1 on X_2 and X_3 . This would be foolish, since X_2 and X_3 are consequence of X_1 , not causes of it. Remember, the data can't tell you what the proper causal ordering is, you have to decide that for yourself.
- You should regress X_3 on X_1 . If you instead regress X_3 on X_2 , you'll get an R^2 , but it will be meaningless (unless, perhaps, you are intentionally using X_2 as a proxy for the unmeasured X_1 .) Such questionable modeling occurs when, say, one attitude is regressed on another, when in reality both attitudes are functions of something else.
- Another way to increase R^2 is to regress a variable on a slightly different operationalization of itself. This might occur if different attitudinal items really measure pretty much the same thing.
- In short, merely increasing R^2 by lengthening the list of regressors is no great achievement unless the role of these variables in an extended model is properly understood and correctly represented.

EVILS OF R^2 (In Cross-population comparisons)

Bivariate R^2 is a function of three quantities:

1. The structural effect of X on Y (β)
2. The variance of the exogenous variable (X)
3. The variance of the disturbance term (u)

More generally, R^2 is a function of the β structural coefficients, the variances of the exogenous variables, and the variances of the disturbance terms. The following examples illustrate this:

EXAMPLES.

1. Exogenous variance differs across populations	
<i>Population 1</i>	<i>Population 2</i>
$\beta = 3$	$\beta = 3$
$V(X) = 4$	$V(X) = 9$
$V(u) = 27$	$V(u) = 27$
$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 4}{9 * 4 + 27} = .57$	$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 9}{9 * 9 + 27} = .75$
The population with the larger exogenous variance will have the larger R^2	

2. Structural effect differs across populations	
<i>Population 1</i>	<i>Population 2</i>
$\beta = 3$	$\beta = 6$
$V(X) = 4$	$V(X) = 4$
$V(u) = 27$	$V(u) = 27$
$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 4}{9 * 4 + 27} = .57$	$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{36 * 4}{36 * 4 + 27} = .84$
The population with the larger structural effect will have the larger R^2	

3. Variance of disturbance differs across populations

<i>Population 1</i>	<i>Population 2</i>
$\beta = 3$	$\beta = 3$
$V(X) = 4$	$V(X) = 4$
$V(u) = 27$	$V(u) = 9$
$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 4}{9 * 4 + 27} = .57$	$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 4}{9 * 4 + 9} = .80$

The population with the smaller DV residual variance will have the larger R^2

4a. Structural effect is smaller in one population but the exogenous variance is greater

<i>Population 1</i>	<i>Population 2</i>
$\beta = 3$	$\beta = 2$
$V(X) = 4$	$V(X) = 16$
$V(u) = 27$	$V(u) = 27$
$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 4}{9 * 4 + 27} = .57$	$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{4 * 16}{4 * 16 + 27} = .70$

In this particular example, even though the effect of X on Y is only 2/3 as large in population 2 as it is in population 1, R^2 winds up being greater in population 2 because population 2 is more variable on X.

4b. Structural effect is smaller in one population but the exogenous variance is greater	
<i>Population 1</i>	<i>Population 2</i>
$\beta = 3$	$\beta = 2$
$V(X) = 4$	$V(X) = 9$
$V(u) = 27$	$V(u) = 27$
$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{9 * 4}{9 * 4 + 27} = .57$	$R^2 = \frac{\beta^2 V(X)}{\beta^2 V(X) + V(u)} = \frac{4 * 9}{4 * 9 + 27} = .57$
In this case, if you just looked at R^2 , you would conclude that the two populations are very similar, when in reality they are quite different. Just because R^2 values are similar does not necessarily mean that populations are similar.	

SUMMARY AND IMPLICATIONS.

- A change in any of the three components (β , the exogenous variable variance, or the variance of the disturbance) can change R^2
- When R^2 differs across populations, the difference could be due to differences in any of the three components. For example, the structural effect and the variances of the disturbance could be the same in all populations, but the variance of the exogenous variable could differ (example 1). And, of course, the effect of X on Y (β) could be greater in one population than the other (example 2). Or, the exogenous variance and the structural effects might be the same in both populations, but the random influences that affect Y (i.e. the disturbance) may be more variable in one population than the other (example 3)
- Therefore, to simply note that R^2 differs across populations is of limited usefulness. It is far more useful to explain why R^2 differs. Is it because the structural effects are greater in one population, e.g. education has a larger effect on the income of men than it does women? Or is it because of differing variabilities in the exogenous variables, e.g. men are more variable in their education than women are? Or is it just because the random influences that affect outcomes are more variable in one population than the other?
- Note further that the structural effects might actually be smaller in one population, yet the R^2 in that population could be larger (example 4a). For example, education might have a smaller effect on the income of men than it does women; but if men are more variable in their education the R^2 for men could be larger. In such a case, it might be tempting to say that, because the male R^2 is higher, education has a greater effect on the income of men than it does women. This is highly misleading though, because in reality the structural effect of education on income is smaller for men than women. That is, R^2 is larger for men, not

because the effect of education is larger for them, but because they are more variable in their levels of education.

- Conversely, two very different populations could have similar R^2 values, obscuring the differences between them (example 4b).
- Similar comments apply for other sorts of R^2 comparisons you might be tempted to make, e.g. changes in R^2 across time, differences in R^2 for different dependent variables.

None of this is to say that R^2 is a meaningless or useless statistic. A low R^2 might well indicate that variables are poorly measured, that important variables have been excluded, or that the model has been mis-specified in other ways (e.g. effects are non-linear or non-additive).

But, this does suggest that R^2 should generally be of only secondary interest to us. If a correctly specified model with well-measured variables produces a small R^2 , then so be it. We should be much more interested in the determinants of R^2 than in R^2 itself. And, if we are going to make comparisons of R^2 , we should make sure we are doing so correctly. Rather than just saying R^2 differs across groups, times, or variables, we should try to explain why it differs (and we should definitely avoid misleading statements, such as those which erroneously imply that a larger R^2 is the result of larger structural effects.)

Appendix: Formula Proofs (Optional). Some of the formulas for R^2 can now be easily proven using the techniques we have recently developed. Take the equation

$$Y = \beta_1 X_1 + \beta_2 X_2 + v$$

If we multiply both sides of the equation by Y and take expectations, we get

$$\begin{aligned} E(Y^2) &= \beta_1 E(X_1 Y) + \beta_2 E(X_2 Y) + E(vY) \\ &= \beta_1 \sigma_{1Y} + \beta_2 \sigma_{2Y} + E(v * [\beta_1 X_1 + \beta_2 X_2 + v]) \\ &= \beta_1 \sigma_{1Y} + \beta_2 \sigma_{2Y} + \beta_1 E(vX_1) + \beta_2 E(vX_2) + E(vv) \\ &= \beta_1 \sigma_{1Y} + \beta_2 \sigma_{2Y} + \sigma_v^2 \\ &= \sum_{k=1}^K \beta_k \sigma_{kY} + \sigma_v^2 \\ &= \text{Explained Variance} + \text{Residual Variance} \end{aligned}$$

where K = the number of IVs in the equation. Of course, in a sample, when variables are standardized, R^2 = explained variance. Hence, for a sample we get the formula

$$R^2 = \sum_{k=1}^K b'_k r_{yk}$$

As an alternative, we can square each side and take expectations. This gives us

$$\begin{aligned} E(Y^2) &= E[(\beta_1 X_1 + \beta_2 X_2 + v)^2] \\ &= E(\beta_1^2 X_1^2 + 2\beta_1 \beta_2 X_1 X_2 + \beta_2^2 X_2^2 + 2\beta_1 v X_1 + 2\beta_2 v X_2 + v^2) \\ &= \beta_1^2 E(X_1^2) + 2\beta_1 \beta_2 E(X_1 X_2) + \beta_2^2 E(X_2^2) + E(v^2) \\ &= \beta_1^2 \sigma_1^2 + 2\beta_1 \beta_2 \sigma_{12} + \beta_2^2 \sigma_2^2 + \sigma_v^2 \\ &= \sum_{i=1}^K \sum_{j=1}^K \beta_i \beta_j \sigma_{ij} + \sigma_v^2 \\ &= \text{Explained variance} + \text{unexplained variance} \end{aligned}$$

In a sample, with standardized variables, for the 2 IV case this gives us

$$R^2 = b_1'^2 + b_2'^2 + 2b_1 b_2' r_{12}$$

and, more generally, we get

$$R^2 = \sum_{i=1}^K \sum_{j=1}^K b_i b_j' r_{ij}$$

(Remember that the correlation of any variable with itself is 1).