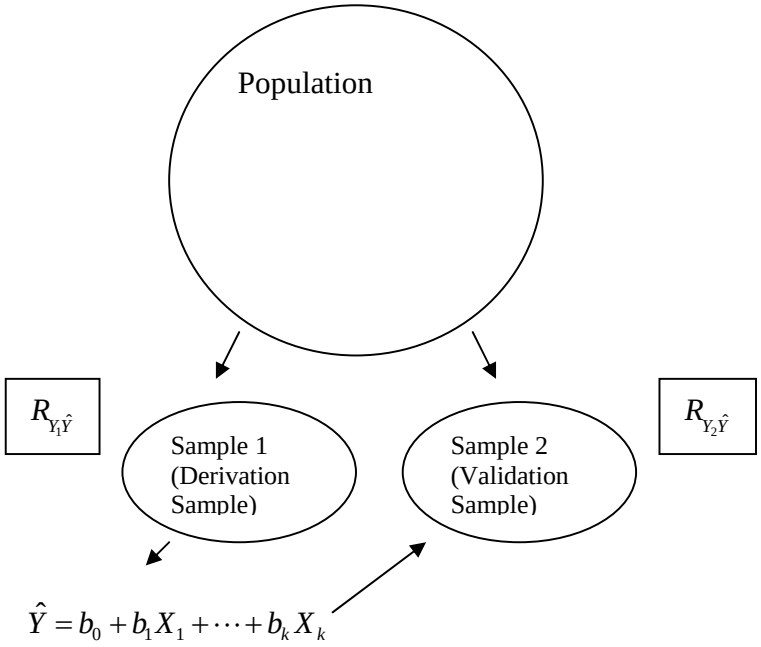
EPRS8550
Model Validation

Figure 1



$R_{Y_1\hat{Y}} > R_{Y_2\hat{Y}}$ Shrinkage of the multiple correlation

- Degree of shrinkage affected by the ratio k / N where k is the number of IVs and N is the sample size. (Some recommend 30 subjects per IV, others N = 400)
- Larger the ratio, larger the overestimation of R.
- Suppose R^2 in the population is 0. The expectation of the sample R^2 is k/(N-1). Consider an extreme case when k = 10 and N = 11. k/(N-1) = 1 (R^2 = 1!!!!!)

1. Adjusted R^2 (R^2_{adj})

Wherry's formula - used in SPSS

$$\hat{\rho}^2 = 1-(1-R^2)\frac{N-1}{N-k-1}$$

where $\hat{\rho}$ is the estimate of ρ, the population multiple correlation coefficient.

Example: R^2_{adj} for three sample sizes and two different R^2 when k = 3

	N = 15	N = 90	N = 150
R^2 = .36	.19	.34	.35
R^2 = .60	.49	.59	.59

Note shrinkage is less if R^2 is larger to start with.

Stein formula

$$\hat{\rho}_c^2 = 1-(1-R^2)(\frac{N-1}{N-k-1})(\frac{N-2}{N-k-2})(\frac{N+1}{N})$$

How effective the sample regression function is in <u>other</u> samples (sample cross validity)

$\hat{\rho}_c^2 < \hat{\rho}^2$

Example:

N = 50, k = 10, R^2 = .50

$\hat{\rho}_c^2 = .191, \hat{\rho}^2 = .372$

2. Cross Validation (Data Splitting)

See Figure 1. In a cross validation study, Sample 1 is called "screening" or "derivation" sample and Sample 2 is called "second" or "late time" or "data splitting" or "validation" sample. $R_{Y_1\hat{Y}}$ is actually compared to $R_{Y_2\hat{Y}}$. If shrinkage is "small" and R^2 is considered meaningful, then, combine two samples and come up with a predication equation for a future use. You can also use a more rigorous approach – double cross validation.

3. PRESS statistic (N validations each based on N - 1observations)

$$PRESS = \sum \hat{e}^2_{(-i)} \qquad R^2_{PRESS} = 1-\frac{PRESS}{(N-1)S_y^2}$$

$\hat{e}^2_{(-i)} = Y_i - \hat{Y}_{(-i)}$ where $\hat{Y}_{(-i)}$ is the predicted Y for Subject i when Subject i was not included to formulate the prediction equation.

An example of misuse of regression analysis

Schutz (1977) “Contingency theory of leadership” (success in administering schools calls for different personality styles depending on the social setting of the school)

Y -- administrative success
X -- 24 personality attributes (k = 24)

Then he showed “ the magnitude of the relationship was greater for subsamples homogeneous with respect to social setting”

total	N = 147 principals
subsamples	$n_1 = 35$, $n_2 = 61$, $n_3 = 36$

Expected $R^2 = k /(N -1) = 24/34 = .706$ when there is NO relationship between Y and X’s!!!!!