Cross-validation:

The one topic that many texts in statistics (and in regression) ignore and that I believe to be crucial to our use of these methods, is cross-validation –

we might discuss this under the rubric of how does one result found for a given sample of data hold up in a new sample

Suppose we look at the squared correlation, R^2, as a measure of how good my *sample* equation is in the *sample* I have – here, we use the squared correlation between Y and $\hat{Y}$

Adjusted R^2 does something similar but more as to how the *population* model does in the *population*

What I am really interested in is how well does the sample equation work more generally – I have optimized the sample equation to the particular data at hand.

In other words, how well does the sample equation do in a new group – the quintessential question of "cross-validation"

How to do it:

a) Get *new* data and use the sample equation to predict Y and get the squared correlation between Y and $\hat{Y}$

call this R^2_{new}

The difference between R^2 and R^2_{new} is called "shrinkage" and measures the drop in how well one can predict with new data (this is not Wherry's formula, irrespective of what some may say) ...

The problem: new data are sometimes "hard to come by"

b) Split the original sample into two; obtain the equation on one part (the "training sample") and test how well it does on on the second (the "test" sample)

Smaller sample sizes in the training sample lead to more unstable equations –

c) *Sample Reuse Methods*: Suppose I break up my samples into K parts; I fit my equation with $K-1$ of the parts together and test on the one part left out. I repeat this process K times, leaving one of the K parts out each time. (This is called K-fold cross-validation.)

At the extreme, if I have n subjects, I could do n-fold cross-validation where I leave one person out at a time;

I predict for this person (say, person i), obtaining $\widehat{Y}_i$, and then obtain the squared correlation between the Y_i's and $\widehat{Y}_i$'s to see how well I cross-validate with a "new" sample.

Each equation I calculate is based on $n - 1$ subjects, so I should have more stability than in (b) –

Jackknife:

An idea similar to the "hold-out-some(one)-at-a-time" is Tukey's Jackknife.

This was devised by Tukey to obtain a confidence interval on a parameter (and indirectly to reduce the bias of an estimator that is not already unbiased)

In Psychology, there is an early discussion of the Jackknife in the Handbook of Social Psychology (Volume II) (Lindzey and Aronson; 1968) by Mosteller and Tukey: Data Analysis — Including Statistics

General approach for the Jackknife:

suppose I have n observations $X_1, \ldots, X_n$ and let θ be an unknown parameter of the population.

We have a way of estimating θ (by, say, $\widehat{\theta}$) –

Group the n observations into t groups of m; thus, $n = tm$:

$\{X_1, \ldots, X_m\}, \ldots, \{X_{(t-1)m+1}, \ldots, X_{tm}\}$

Let $\widehat{\theta}_{-0}$ be the estimate based on all groups;

Let $\widehat{\theta}_{-i}$ be the estimate based on all groups except the i^{th}

Define new estimates of θ, called "pseudo-values" as follows:

$\widehat{\theta}_{*i} = t\widehat{\theta}_{-0} - (t-1)\widehat{\theta}_{-i}$, for $i = 1, \ldots, t$

The Jackknife estimate of θ is the mean of the pseudo-values:

$\widehat{\theta}_{*\cdot} = \sum_{i=1}^{t} \frac{\widehat{\theta}_{*i}}{t}$

An estimate of its standard error is

$s_{\widehat{\theta}_{*\cdot}} = [\sum_{i=1}^{t} \frac{(\widehat{\theta}_{*i} - \widehat{\theta}_{*\cdot})^2}{t(t-1)}]^{1/2}$

Approximate confidence interval:

$\widehat{\theta}_{*\cdot} \pm s_{\widehat{\theta}_{*\cdot}} t_{\frac{\alpha}{2},t-1}$

We act as if the t pseudo-values $\widehat{\theta}_{*1}, \ldots, \widehat{\theta}_{*t}$ are independent and identically distributed observations.

We also reduce some bias in estimation if the original estimate we used was biased.

An example:

suppose I want to estimate μ based on $X_1, \ldots, X_n$

Choose $t = n$

$\widehat{\theta}_{-0} = \frac{1}{n} \sum_{j=1}^{n} X_j$

$\widehat{\theta}_{-i} = \frac{1}{n-1} \sum_{j=1, i \neq j}^{n} X_j$

$\widehat{\theta}_{*i} = n(\frac{1}{n} \sum_{j=1}^{n} X_j) - (n-1)(\frac{1}{n-1} \sum_{j=1, i \neq j}^{n} X_j) =$

X_i

Thus, $\widehat{\theta}_{*\cdot} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$

$$s_{\widehat{\theta}_{*\cdot}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2} =$$

$\sqrt{s_X^2/n}$, where s_X^2 is an unbiased estimate of σ^2

Confidence interval:

$$\bar{X} \pm (\sqrt{s_X^2/n})\ t_{\frac{\alpha}{2},t-1}$$

The Bootstrap:

Population ("Theory World"): the pair of random variables X and Y are, say, bivariate normal

Sample ("Data World"): n pairs of independent and identically distributed observations on (X, Y):

$(X_1, Y_1), \ldots, (X_n, Y_n)$; these could be used to give r_{XY} as an estimate of ρ_{XY}

Now, make Data World the Theory World Population:

$(X_1, Y_1), \ldots, (X_n, Y_n)$, and each occurs with probability $\frac{1}{n}$

Sample this Theory World Population (with replacement) to get one "bootstrap" sample (with possible repeats):

$(X_1^{'}, Y_1^{'}), \ldots, (X_{n^{'}}^{'}, Y_{n^{'}}^{'})$ (usually, n equals $n^{'}$)

Get B bootstrap samples and compute the correlation for each: $r_{XY}^{(1)}, \ldots, r_{XY}^{(B)}$

This last distribution could be used, for example, to obtain a confidence interval on ρ_{XY}

Permutation tests for correlation measures:

We start at the same place as for the Bootstrap:

Population ("Theory World"): the pair of random variables X and Y are, say, bivariate normal

Sample ("Data World"): n pairs of independent and identically distributed observations on (X, Y):

$(X_1, Y_1), \ldots, (X_n, Y_n)$; these could be used to give r_{XY} as an estimate of ρ_{XY}

Now, to test H_o : X and Y are statistically independent.

Under H_o, the X's and Y's are matched at random; so, assuming (without loss of generality) that we fix the X's, all $n!$ permutations of the Y's against the X's are equally likely to occur.

We can calculate a correlation for each of these $n!$ permutations and graph:

the distribution is symmetric and unimodal at zero; the range along the horizontal axis obviously goes from -1 to $+1$

$p-$value (one-tailed) = number of correlations as or larger than the observed correlation/$n!$

Also, as an approximation, $r_{XY} \sim N(0, \frac{1}{n-1})$;

Thus, the standard error is close to $\frac{1}{\sqrt{n}}$; this might be useful for quick "back-of-the-envelope" calculations