regress

syntax: `[params, const] = regress(dep, indep )`
`[params, const, r2] = regress(dep, indep )`
`[params, const, r2, resids] = regress(dep, indep )`

purpose: Carries out multiple linear regression to find the best fit of `dep` as a linear function of `indep`.

examples: Suppose we make a series of pairs of measurements. For example, in the stork-and-babies data

```
storks = [1920 1700 1090  990 1030  995  930];
babies = [1.04 0.78 0.62 0.58 0.57 0.58 0.62]*1000000;
```

We suspect that there is a straight-line relationship between the number of storks and the number of babies, and we want to find the parameters a and b in the linear formula

$$\text{babies} = b + a * \text{storks}$$

We can do this with
≫ `[a,b] = regression(babies', storks')`
There are two things to note about this:

1. The first returned argument is the slope a in the linear formula, the second is the y-intercept term b.
2. The ' operator has been used to transpose the data, which is in row vector form, into column data.

Also note that we multiplied the baby data by 1 million so that the units are in "babies" rather than "millions of babies." This is just to make the interpretation a little easier.

The result of the regression is $a = 407.7$ and $b = 180110$. This can be interpreted as meaning that each stork accounts for 407.7 babies per year. If there were no storks, the annual number of births would be 180,110.

The syntax
≫ `[a,b,r2] = regression(babies', storks')`
causes the r^2 measure of goodness of fit to be calculated. In this case, $r^2 = .89$ which indicates a strong fit.

The syntax
≫ `[a,b,r2,resids] = regression(babies', storks')`
will give the "residuals" from the fitted line.

When there is more than one independent variable, `regression` will perform multiple regression. This is done by packaging the independent variables into a matrix, with one dependent variable per column. In the stork-and-baby data, we also have the year in which the measurement was made.

```
years  = [1965 1971 1974 1977 1978 1979 1980];
```

We might want to take what seems to be a trend to decreases in the number of births into account at the same time as we find how many babies each stork brings each year. To do this, we would be interested in the equation:

$$\text{babies} = b + a1 * \text{storks} + a1 * \text{years}$$

The parameters $a1$, $a2$, and b can be found using `regression`:
≫ `[a,b,r2] = regression(babies', [storks', years'])`
Now, `a` consists of two values, `[ 250 -12450]`, one for each of the two independent variables used in the regression. The constant term, `b` is 24,963,000. The interpretation is that the number of births falls by 12450 each year, and that each additional member of the stork population accounts for 250 births each year. If there were no storks, and set the year to be 0 in the equation (this extrapolates the linear function backwards to the year zero) then the number of births per year would be 24,963,000. The goodness of fit is indicated by r^2, which is now 0.90, better than when just `storks` was used. (If we carry out the regression using just the `years` data, we would still find quite a good fit, with $r^2 = 0.87$, a =-30,234 and b =60,392,000.)

This is, of course, a frivolous example. It serves, however, to illustrate several aspects of linear regression: the existence of a fit with a large r^2 does not imply causality; adding in more independent variables will tend to increase r^2 (even when the independent variable is unrelated to the dependent variable); in multiple regression, the independent variables may be linearly related themselves, so including more than one variable can lessen the impact of the variables individually; a linear model may not be appropriate; and extrapolating a linear model (e.g., finding the number of births in the year zero) can give completely misleading results.

Because `regress` is intended for resampling operations, conventional significance information (t-values, F-values, etc.) is not provided.

See also: REGRESSIONVERBOSE which prints out the results of the regression.

CORR computes r^2 for two variables. The function `regress` in the MATLAB statistics toolbox carries out multiple regression in a similar fashion (but without taking into account the constant term) and returns some traditional measures of significance of the fitted parameters.