Normal Correlation Models:

Suppose (X, Y) is bivariate normal; then

$Y|X = x \sim$

$N(\mu_Y + \rho_{YX}(\frac{\sigma_Y}{\sigma_X})(x - \mu_X), \sigma_Y^2(1 - \rho_{YX}^2)) =$

$N((\mu_Y - \rho_{YX}(\frac{\sigma_Y}{\sigma_X})\mu_X) + \rho_{YX}(\frac{\sigma_Y}{\sigma_X})x,$

$\sigma_Y^2(1 - \rho_{YX}^2)) =$

$N(\beta_0 + \beta_1 x, \sigma^2)$

We test $H_o : \rho_{YX} = 0$ (or, $H_o : \beta_1 = 0$) with the usual t-tests

We use Fisher's Z-transformation to get a confidence interval on ρ_{YX}

Here, the correlation r_{YX} estimates a parameter; otherwise, it really is only a descriptive statistic

Now, in the more general case where

$(Y, X_1, \ldots, X_{p-1})$ is multivariate normal:

$(Y|X_1 = x_1, \ldots, X_{p-1} = x_{p-1}) \sim$

$N(\mu_Y + \beta_1(x_1 - \mu_1) + \cdots + \beta_{p-1}(x_{p-1} - \mu_{p-1}),$

$\sigma_Y^2(1 - \rho^2_{Y \cdot 12\ldots(p-1)})) =$

$N(\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}, \sigma^2)$

where $\beta_0 = \mu_Y - \beta_1\mu_1 - \cdots - \beta_{p-1}\mu_{p-1}$ and

$\sigma^2 = \sigma_Y^2(1 - \rho^2_{Y \cdot 12\ldots(p-1)})$

The population squared multiple correlation is $\rho^2_{Y \cdot 12\ldots(p-1)}$

We can show several interesting things:

$$= \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{(p-1)\times 1} =$$

$$\begin{bmatrix} V(X_1) & \cdots & Cov(X_1, X_{p-1}) \\ \vdots & & \vdots \\ Cov(X_{p-1}, X_1) & \cdots & V(X_{p-1}) \end{bmatrix}^{-1}_{(p-1)\times(p-1)} \times$$

$$\begin{bmatrix} \vdots \\ cov(X_i, Y) \\ \vdots \end{bmatrix}_{(p-1)\times 1}$$

$\rho^2_{Y\cdot 12\ldots(p-1)}$ = (the population squared multiple correlation coefficient)

$$\begin{bmatrix} \vdots \\ cov(X_i, Y) \\ \vdots \end{bmatrix}' \times$$

$$\begin{bmatrix} V(X_1) & \cdots & Cov(X_1, X_{p-1}) \\ \vdots & & \vdots \\ Cov(X_{p-1}, X_1) & \cdots & V(X_{p-1}) \end{bmatrix}^{-1} \times$$

$$\begin{bmatrix} \vdots \\ cov(X_i, Y) \\ \vdots \end{bmatrix} / \sigma^2$$

This is also the squared correlation between Y and $\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$

Again, estimates can be put in and we get a sample squared multiple correlation coefficient, $R^2_{Y \cdot 1 \ldots (p-1)}$

To test $H_o : \rho^2_{Y \cdot 12 \ldots (p-1)} = 0$

$$\frac{R^2_{Y \cdot 12 \ldots (p-1)}}{1 - R^2_{Y \cdot 12 \ldots (p-1)}} (\frac{n-p}{p-1}) \sim F_{p-1, n-p}$$

Numerically, this is the same as we did before.

When we standardize all our variables to mean zero and variance one, the covariances become correlations.

All of multiple regression can be done on correlations alone; all the regression coefficients are standardized and $\beta_0 = 0$

What happens if all the independent variables are uncorrelated?

what are the regression coefficients then?

Partial Correlation:

Suppose I have three variables Y, X_1, X_2

I would like to have some way of assessing the following question:

What is the relation between Y and X_1 after you "control for X_2"; or "hold X_2 constant'; or "get rid of the effect of X_2"

Approach this question in the following way:

$\widehat{Y}_i = b_0 + b_1 X_{i2}$

$\widehat{X}_{i1} = b_0^* + b_1^* X_{i2}$

Look at the residuals from these regressions on X_2: $Y_i - \widehat{Y}_i$ and $X_{i1} - \widehat{X}_{i1}$

These are "free" of the X_2 variable –

If I correlate the residuals, $Y_i - \widehat{Y}_i$ and $X_{i1} - \widehat{X}_{i1}$, I have the partial correlation of Y and X_1 "holding X_2 constant"

This is denoted by $r_{YX_1 \cdot X_2}$ or $r_{Y1\cdot 2}$

If I square it, I get the "coefficient of partial determination": $r^2_{Y1\cdot 2}$

Obviously, I can do this more generally as well, e.g., define $r_{Y1\cdot 234\ldots(p-1)}$

In fact, there are formulas for these: e.g.,

$$r_{Y1\cdot 2} = \frac{r_{Y1} - r_{12}r_{Y2}}{\sqrt{(1 - r^2_{12})(1 - r^2_{Y2})}}$$

$$r_{Y1\cdot23} = \frac{r_{Y1\cdot3} - r_{12\cdot3}r_{Y2\cdot3}}{\sqrt{(1 - r^2_{12\cdot3})(1 - r^2_{Y2\cdot3})}}$$

There are interesting connections with the multiple regression model:

$$r^2_{Y1\cdot23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)} = 1 - \frac{SSE(X_1, X_2, X_3)}{SSE(X_2, X_3)}$$

$$r^2_{Y1\cdot2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = 1 - \frac{SSE(X_1, X_2)}{SSE(X_2)}$$

In testing partial correlations or putting confidence intervals on (using Fisher Z-transformations), treat partial correlations as regular correlations but reduce the degrees of freedom by 1 for each variable "held constant"

Some other connections with multiple regression:

Suppose I have two independent variables, X_1 and X_2, in my model

$$R^2 = 1 - \frac{SSE(X_1, X_2)}{SSTO} = \frac{SSR(X_1, X_2)}{SSTO} =$$

$$\frac{SSR(X_2)}{SSTO} + \frac{SSR(X_1|X_2)}{SSTO} =$$

$$r^2_{Y2} + r^2_{Y(1\cdot 2)}$$

the latter term is called the squared part or semipartial correlation

Also, we can give some formulas:

$$r_{Y(1\cdot 2)} = \frac{r_{Y1} - r_{12}r_{Y2}}{\sqrt{(1 - r_{12}^2)}} \leq r_{Y1\cdot 2}$$

This is also the correlation between Y and $X_{i1} - \hat{X}_{i1}$, where

$\hat{X}_{i1} = b_0^* + b_1^* X_{i2}$

Selection of Independent Variables:

The general question: starting with a (large) set of variables, can we find a smaller subset that does well

---

In multivariate analysis, it is important to remember that there is systematic covariation possible among the variables, and this has a number of implications for how we proceed.

Automated analysis methods that search through collections of independent variables to locate the "best" regression equations (for example, by forward selection, backward elimination, or the hybrid of stepwise regression) are among the most misused statistical methods available in software packages.

They offer a false promise of blind theory building without user intervention, but the incongruities present in their use are just too great for this to be a reasonable strategy of data analysis:

(a) one does not necessarily end up with the "best" prediction equations for a given number of variables;

(b) different implementations of the process don't necessarily end up with the same equations;

(c) given that a system of interrelated variables is present, the variables not selected cannot be said to be unimportant;

(d) the order in which variables enter or leave in the process of building the equation does not necessarily reflect their importance;

(e) all of the attendant significance testing and confidence interval construction methods become completely inappropriate.

Several methods, such as the use of Mallow's C_p statistic for "all possible subsets (of the independent variables) regression," have some possible mitigating effects on the heuristic nature of the blind methods of stepwise regression.

They offer a process of screening all possible equations to find the better ones, with compensation for the differing numbers of parameters that need to be fit.

Although these search strategies offer a justifiable mechanism for finding the "best" according to ability to predict a dependent measure, they are somewhat at cross-purposes for how multiple regression is typically used in the behavioral sciences.

What is important is the structure among the variables as reflected by the regression, and not so much squeezing the very last bit of variance accounted for from our data.

More pointedly, if we find a "best" equation with fewer than the maximum number of available independent variables present, and we cannot say that those not chosen are less important than those that are, then what is the point?

The general problem of model adequacy:

We made the *assumption* of a linear regression model, but we also should be concerned with how well the model represents the data.

One way was formally in our lack-of-fit test when we had appropriate repeats.

More informally we look at the residuals once one has fitted the least squares line; hopefully this might lead to suggestions for obtaining a better model.

Rather subjective approach that relies heavily on looking at graphs, and in particular, whether the estimated residuals, $e_i = Y_i - \hat{Y}_i$, look more or less normal with constant variance.

For example, does a nonlinearity of residuals suggest we should have included other variables? (e.g., X^2)

Do the residuals show a nonconstant variance that might be helped by transformations of the variables or the use of weighted least-squares?

Do the residuals show a pattern if plotted against another variable, such as time of observation?

Are there obvious outliers among the residuals that should be explained?

We might also plot other variables directly in the residual plot (e.g., we might label the residuals as to male or female by using blue or pink)