The Correlational Model: Correlation problems in the population:

We observe pairs of numbers (X, Y) and *both* X and Y are random variables.

Example, X may denote some "I.Q." score and Y an achievement score obtained for one subject –

Suppose we standardize these two random variables: $\frac{X-\mu_X}{\sigma_X}$ and $\frac{Y-\mu_Y}{\sigma_Y}$

Before, in descriptive statistics we took the sum of the multiplied z-scores and divided by N:

$\frac{\sum Z_X Z_Y}{N} = r_{XY}$.

This now becomes $E[(\frac{X-\mu_X}{\sigma_X})(\frac{Y-\mu_Y}{\sigma_Y})] = \rho_{XY}$,

the population correlation coefficient.

$E[(\frac{X-\mu_X}{\sigma_X})(\frac{Y-\mu_Y}{\sigma_Y})] =$

$(\frac{1}{\sigma_X \sigma_Y})E[(X - \mu_X)(Y - \mu_Y)] =$

Covariance$(X, Y)/(\sigma_X \sigma_Y) =$

Cov$(X, Y)/(\sigma_X \sigma_Y) = \rho_{XY}$

In the sample:

$r_{XY} = (\frac{1}{N}) \sum_{i=1}^{N} (\frac{X_i - M_X}{S_X})(\frac{Y_i - M_Y}{S_Y}) =$

$\frac{1}{S_X S_Y}[(\frac{1}{N}) \sum_{i=1}^{N} (X_i - M_X)(Y_i - M_Y)] =$

$\frac{1}{S_X S_Y}$ times the sample covariance between X and Y.

Fact: if X and Y are independent, then cov$(X, Y) =$ 0 (and also, the correlation r_{XY} is zero)

This is not true the other way around unless we make a more stringent assumption that X and Y have a bivariate normal distribution – then cov$(X, Y) = 0$ (or $r_{XY} = 0$) if and only if X and Y are statistically independent.

If X and Y are bivariate normal, then $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$ and when taken together X and Y have a certain joint distribution (it looks like a three-dimensional bell sitting on the X and Y plane.

If you sever the bell horizontally at a certain height, an ellipse is drawn out whose orientation and "tightness" (or elongation) are determined by ρ_{XY} – a value of plus or minus 1.0 gives a line (a degenerate ellipse); a value of 0.0 gives a circle.

There are a number of facts about correlation problems that are directly pertinent to descriptive aspects discussed previously.

1) $-1 \leq \rho_{XY} \leq +1$

2) If X and Y are bivariate normal, then knowing the value for X gives you some information about Y:

$E(Y|X) = \rho_{XY}(\frac{\sigma_Y}{\sigma_X})(X - \mu_X) + \mu_Y$; this is called "the regression of Y on X" (the conditional expectation of Y given X)

3) $\beta_{Y \cdot X} = \rho_{XY}(\frac{\sigma_Y}{\sigma_X})$ is the population regression coefficient

4) $\sigma^2_{Y \cdot X} = \sigma^2_Y(1 - \rho^2_{XY})$ is the variance of the random variable Y given that you know an X value

5) $\sigma_{Y \cdot X} = \sigma_Y \sqrt{(1 - \rho^2_{XY})}$ is the standard error of estimate

6) the proportion of variance accounted for by linear regression: $\frac{\sigma^2_Y - \sigma^2_{Y \cdot X}}{\sigma^2_Y} = \rho^2_{XY}$

7) our best estimate of the population regression equation:

$\rho_{XY}(\frac{\sigma_Y}{\sigma_X})(X - \mu_X) + \mu_Y =$

$\beta_{Y\cdot X}(X - \mu_X) + \mu_Y$ is

$b_{Y\cdot X}(X - M_X) + M_Y$, the descriptive equation

8) finally, an unbiased estimate of $\sigma^2_{Y\cdot X}$ is $\frac{N}{N-2}S^2_{Y\cdot X}$, i.e.,

$\hat{\sigma}^2_{Y\cdot X} = \frac{N}{N-2}S^2_Y(1 - r^2_{XY})$

Inference procedures in the normal (correlational) model:

First, note that $E(r_{XY}) \neq \rho_{XY}$, i.e., the sample correlation is not an unbiased estimate of the population correlation. There are ways of correcting the sample correlation to make it unbiased but I have never seen it done in practice.

To test $H_o : \rho_{XY} = 0$, use $t = \frac{r_{XY}\sqrt{N-2}}{\sqrt{1-r^2_{XY}}}$ with $N - 2$ degrees of freedom.

The same test can be used for $H_o : \beta_{Y \cdot X} = 0$

The sampling distribution of the sample correlation, r_{XY}, has a difficult form (when the population correlation is not zero) where the sampling variance is a function of the population correlation, ρ_{XY}.

To work around this problem for testing a nonzero value of a correlation or constructing a confidence interval on ρ_{XY}, Fisher's r to Z transformation is used:

$Z = \frac{1}{2}\log_e(\frac{1+r_{XY}}{1-r_{XY}})$, which is the inverse hyperbolic tangent function ($\tanh^{-1}$) –

We know that approximately $Z \sim N(E(Z), \frac{1}{N-3})$ where $E(Z) = \frac{1}{2}\log_e(\frac{1+\rho_{XY}}{1-\rho_{XY}})$; $Var(Z) = \frac{1}{N-3}$ as indicated, and does not include the population correlation ρ_{XY} (Z is called a variance-stabilizing transformation).

To get a confidence interval on ρ_{XY}, first get a confidence interval on $E(Z) = \frac{1}{2}\log_e(\frac{1+\rho_{XY}}{1-\rho_{XY}})$ as $Z \pm z_{.025}\sqrt{\frac{1}{N-3}}$ and work the tables backward (here, $z_{.025} = 1.96$)

There is a nice review paper by James Steiger about “tests for comparing elements of a correlation matrix” (Psychological Bulletin, 1980, 87, 245-251). Here’s the abstract:

In a variety of situations in psychological research, it is desirable to be able to make statistical comparisons between correlation coefficients measured on the same individuals.

For example, an experimenter may wish to assess whether two predictors correlate equally with a criterion variable.

In another situation, the experimenter may wish to test the hypothesis that an entire matrix of correlations has remained stable over time.

The present article reviews the literature on such tests, points out some statistics that should be avoided, and presents a variety of techniques that can be used safely with medium to large samples.

Several illustrative numerical examples are provided.

————————

An example of the approach(es) used by Steiger:

Suppose $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent and estimate θ_1 and θ_2 respectively. I know $\mathrm{Var}(\widehat{\theta}_1)$ and $\mathrm{Var}(\widehat{\theta}_2)$ and that $\widehat{\theta}_1 \sim N(\theta_1, \mathrm{Var}(\widehat{\theta}_1))$ and $\widehat{\theta}_2 \sim N(\theta_2, \mathrm{Var}(\widehat{\theta}_2))$

Thus, $\frac{(\widehat{\theta}_1 - \widehat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\mathrm{Var}(\widehat{\theta}_1) + \mathrm{Var}(\widehat{\theta}_2)}} \sim N(0,1)$

Using this result we can test $H_o : \theta_1 = \theta_2$ or put a confidence interval on $\theta_1 - \theta_2$. If we know the covariance between the estimates, we could also do the same for estimates that are not independent.

Linear Regression Model:

We essentially would like to get to the same results as we did for the bivariate normal model

Assume the following theoretical model that relates Y and X:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Y_i is the dependent variable, the criterion variable, the response variable, or if you are a (pretentious) economist, the endogenous variable

X_i is assumed fixed (i.e., not a random variable); called the independent variable, or the predictor variable, or the exogenous variable

ϵ is a random variable for error and induces Y_i to also be a random variable; the errors ϵ_i are n observations on $\epsilon \sim (0, \sigma^2)$;

could assume that a) the errors are independent so $E(\epsilon_i \epsilon_j) = E(\epsilon_i)E(\epsilon_j) = 0$, or b) that the errors have zero covariance or are uncorrelated $-$

From these assumptions, $E(Y_i) = \beta_0 + \beta_1 X_i$ and $V(Y_i) = \sigma^2$

These are very much like the forms in the normal correlational model for $E(Y|X = x)$ and $V(Y|X = x)$

Generically, we can write $Y = \beta_0 + \beta_1 X + \epsilon$.

The $Y_1, \ldots, Y_n$ are independent (uncorrelated) because $\epsilon_1, \ldots, \epsilon_n$ are independent (uncorrelated)

In both cases (models), for a given value on X, there is a probability distribution for Y – the mean of these distributions vary with X (i.e., the regression function of Y on X);

the variance of these distributions do not

The simple linear regression model – simple because we have only one independent variable; linear in the parameters (they don't sit as exponents, for example); linear in the independent variable (we don't have the squared independent variable in the equation)

Could also phrase as a deviation model:

$Y = \beta_0^* + \beta_1(X - \bar{X}) + \epsilon$, where $\beta_0^* = \beta_0 + \beta_1\bar{X}$

In the simple linear regression model, how do you estimate β_0 and β_1:

Least-squares: $\widehat{\beta}_0 \equiv b_0 = M_Y - \frac{r_{XY}S_Y}{S_X}M_X = (\bar{Y} - \frac{r_{XY}S_Y}{S_X}\bar{X})$

$\widehat{\beta}_1 \equiv b_1 = \frac{r_{XY}S_Y}{S_X}$

These are numerically the same as before in the correlational model.

Properties: (Gauss-Markov theorem)

a) Unbiased: $E(b_0) = \beta_0$ and $E(b_1) = \beta_1$

b) Minimum variance among all unbiased linear estimates:

$b_1 = \sum_{i=1}^{n} k_i Y_i$, where $k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$

$b_0 = \sum_{i=1}^{n} l_i Y_i$, where $l_i = (\frac{1}{n} - \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2})$

Simple Linear Regression Model:

Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = E(Y_i) + \epsilon_i$, where $\epsilon \sim (0, \sigma^2)$

Fitted: $Y_i = b_0 + b_1 X_i + e_i = \widehat{Y}_i + e_i$,

where $\widehat{Y}_i$ $(= b_0 + b_1 X_i)$ is called the "fitted value"

$E(b_0 + b_1 X_i) = \beta_0 + \beta_1 X_i$,

so $\widehat{Y}_i$ is an unbiased estimate of $E(Y_i) = \beta_0 + \beta_1 X_i$;

also, minimum variance in the class of unbiased linear estimators

Residual:

Model: $Y_i - E(Y_i) = \epsilon_i$

Fitted: $Y_i - \widehat{Y}_i = e_i$

One can't say, however, that e_i is an unbiased estimator of ϵ_i since ϵ_i is actually a random variable:

$E(e_i) = E(Y_i - \widehat{Y}_i) = E(Y_i) - E(\widehat{Y}_i) = \beta_0 + \beta_1 X_i - (\beta_0 + \beta_1 X_i) = 0$

Some properties of the fitted regression line:

1) $\sum_{i=1}^{n} e_i = 0$

2) $\sum_{i=1}^{n} e_i^2$ is at a minimum because of the least squares line

3) Because of (1), $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$

4) $\sum_{i=1}^{n} X_i e_i = 0$

5) $\sum_{i=1}^{n} \widehat{Y}_i e_i = 0$

6) Because $\widehat{Y}_i = \bar{Y} + \frac{r_{XY} S_Y}{S_X}(X_i - \bar{X})$,

the regression line goes through $(\bar{X}, \bar{Y})$ (if $X_i = \bar{X}$, then $\widehat{Y}_i = \bar{Y}$)

Estimation of the Error Variance:

As we said in the correlational model, an unbiased estimate of $\sigma^2_{Y\cdot X}$ is $\frac{N}{N-2}S^2_{Y\cdot X}$,

i.e., $\hat{\sigma}^2_{Y\cdot X} = \frac{N}{N-2}S^2_Y(1 - r^2_{XY})$

Using n for N, this latter term is

$\frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{Y}_i) = \frac{1}{n-2}\sum_{i=1}^n e_i^2$

The term $\sum_{i=1}^n e_i^2$ is called the "sum of squared errors" (SSE) (or the "residual sum of squares" or the "error sum of squares");

$\frac{1}{n-2}\sum_{i=1}^n e_i^2$ is the "mean square error" (MSE) (or the "error mean square" or the "residual mean square")

Normal error model:

If we make the further assumption that

$\epsilon_i \sim N(0, \sigma^2)$, then $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.

The parallel to the normal correlational model is then complete (uncorrelatedness implies statistical independence).

The parameter estimates, b_0 and b_1, are then maximum likelihood estimates; the maximum likelihood estimate for σ^2 is the biased estimate with a division by n rather than $n-2$

These estimates are consistent, sufficient, and minimum variance unbiased (period, without qualification as to being linear estimates)

Inference Using the Regression Model:

Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\beta_0 + \beta_1 X_i$ is a constant, and $\epsilon_i \sim N(0, \sigma^2)$

Three basic sampling distribution results:

1) $b_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2})$

so, $(b_1 - \beta_1)/\sqrt{\frac{\text{MSE}}{\sum(X_i - \bar{X})^2}} \sim t_{n-2}$

2) $b_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}))$

so, $(b_0 - \beta_0)/\sqrt{\text{MSE}(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2})} \sim t_{n-2}$

3) $\widehat{Y}_h = b_0 + b_1 X_h$ estimates $\beta_0 + \beta_1 X_h$ which is $E(Y_h)$ when the independent variable is X_h

$\widehat{Y}_h \sim N(\beta_0 + \beta_1 X_h, \sigma^2(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}))$

so, $(\widehat{Y}_h - (\beta_0 + \beta_1 X_h))/\sqrt{\text{MSE}(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2})}$

$\sim t_{n-2}$

Note the effect of the various terms in the variance.

Prediction Intervals:

Suppose I would like to predict where a *new* observation, Y_h, will be (and not just give a confidence interval on $E(Y_h) = \beta_0 + \beta_0 X_h$)

Because Y_h and $\widehat{Y}_h$ are independent,

$$\widehat{Y}_h - Y_h \sim N(0, \sigma^2(1 + \tfrac{1}{n} + \tfrac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}))$$

$$(\widehat{Y}_h - Y_h)/\sqrt{\mathrm{MSE}(1 + \tfrac{1}{n} + \tfrac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2})} \sim t_{n-2}$$

This latter expression leads to the prediction (not a confidence) interval for Y_h as

$$\widehat{Y}_h \pm (\mathrm{t-value})\sqrt{\mathrm{MSE}(1 + \tfrac{1}{n} + \tfrac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2})}$$

The Analysis-of-Variance (ANOVA) Approach to Regression:

Trivial result: $Y_i - \bar{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y})$

$\sum(Y_i - \bar{Y})^2 = \sum[(Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y})]^2$

After some algebra, this reduces to

$\sum(Y_i - \widehat{Y}_i)^2 + \sum(\widehat{Y}_i - \bar{Y})^2$

Or, the sum of squares total (SSTO) (i.e., the amount of "stuff" we would like to explain) is additively decomposed into the sum of squared error (SSE) (i.e.., the amount of unexplained "stuff") and the sum of squares regression (SSR) (i.e., the amount of explained "stuff") (so, in summary, SSTO = SSE + SSR)

Remember that the mean of the $\widehat{Y}_i$'s is $\bar{Y}$

The degrees of freedom for SSTO also decomposes: $n-1$ (SSTO) = ($n-2$) (SSE)+1 (SSR)

A convenient computational formula:

$\text{SSR} = b_1^2 \sum (X_i - \bar{X})^2$

Mean Square Regression (MSR) = SSR/1

Mean Square Error (MSE) = SSE/$(n-2)$

$E(\text{MSE}) = \sigma^2$

$E(\text{MSR}) = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$

Remember when we are dealing with "variance" estimators in a ratio (i.e., $\frac{\text{s}^2}{\sigma^2}$), if the top estimator estimates the bottom parameter unbiasedly, we have $\sim \frac{\chi^2_\nu}{\nu}$

Thus, $\frac{\text{MSE}}{\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2}$ and under $H_o : \beta_1 = 0$,

$\frac{\text{MSR}}{\sigma^2} \sim \frac{\chi^2_1}{1}$, and these two random variables are independent.

Thus, under H_o, $\frac{\text{MSR}/\sigma^2}{\text{MSE}/\sigma^2} = \frac{\text{MSR}}{\text{MSE}}$

$$\sim \frac{\frac{\chi^2_1}{1}}{\frac{\chi^2_{n-2}}{n-2}} = F_{1,n-2}$$

Thus, we have the familiar ANOVA table:

Source (of variation)	df	SS	MS	F
Regression	1	SSR	MSR	MSR/MSE
Error	$n-2$	SSE	MSE	
Total	$n-1$	SSTO		

Test for lack of linear regression:

Suppose I have repeat observations at c levels of X: $X_1, \ldots, X_c$

$Y_{11}, \ldots Y_{n_1 1}$, $Y_{12}, \ldots Y_{n_2 2}$, $\ldots$, $Y_{1c}, \ldots Y_{n_c 1}$

Let $\bar{Y}_1, \ldots, \bar{Y}_c$ denote the means on Y for the c levels of X for which we have repeats

$\sum_{j=1}^{c} n_j = n$ and $\bar{Y}$ is the grand mean

We take another look at the decomposition of SSTO (sum of squares total) into SSE (sum of squares error) and SSR (sum of squares regression

$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$

SSTO $= \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y})^2$

SSR $= \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\hat{Y}_{ij}-\bar{Y})^2$

(but since $\widehat{Y_{ij}} = \widehat{Y_j}$) this is equal to

$\sum_{j=1}^{c}\sum_{i=1}^{n_j}(\hat{Y}_j-\bar{Y})^2 = \sum_{j=1}^{c}(\hat{Y}_j-\bar{Y})^2\sum_{i=1}^{n_j}1 =$

$\sum_{j=1}^{c}n_j(\hat{Y}_j-\bar{Y})^2$

SSE $= \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \widehat{Y}_{ij})^2 =$

$\sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \widehat{Y}_j)^2 =$

$\sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$ (sum of squares for pure error (SSPE)) $+$

$\sum_{j=1}^{c} n_j (\bar{Y}_j - \widehat{Y}_j)^2$ (sum of squares for lack of fit (SSLF))

SSE has $n-2$ degrees of freedom, split into $n-c$ ($= n_1 - 1 + \cdots + n_c - 1$) for SSPE; and $n-2-(n-c) = c-2$ for SSLF

MSPE = SSPE/$(n-c)$

MSLF = SSLF/$(c-2)$

$E(\text{MSPE}) = \sigma^2$ (always)

$E(\text{MSLF}) = \sigma^2 + \frac{\sum_{j=1}^{c} n_j (E(Y_i) - (\beta_0 + \beta_1 X_j))^2}{c-2}$

Thus, under $H_o : E(Y_i) = \beta_0 + \beta_1 X_j$,

MSLF/MSPE $\sim F_{c-2,n-c}$

—————————

Run SYSTAT, R, and Matlab demos –