

General Regression Formulae

Single Predictor Standardized Parameter Model: $Z_{Yi} = \beta Z_{Xi} + \varepsilon_i$

Single Predictor Standardized Statistical Model: $\hat{Z}_{Yi} = \hat{\beta} Z_{Xi}$

Estimate of Beta (Beta-hat): $\hat{\beta} = r_{YX}$ (1)

Standard error of estimate: $s_{Zy.Zx} = \sqrt{1 - r_{YX}^2}$ (2)

Standard error of Beta: $Se_{\beta} = \sqrt{\frac{(1 - r_{YX}^2)}{(N-2)}}$ (3)

There are two identical null hypotheses: $H_0: \beta = 0$ and $H_0: \rho = 0$

Both are tested with a t -statistic with $(N - 2)$ degrees of freedom (df) which can be computed two ways.

$$t_{(N-2)} = r \sqrt{\frac{(N-2)}{1 - r_{YX}^2}} \quad (4) \quad \text{and} \quad t_{(N-2)} = \frac{\hat{\beta} - 0}{Se_{\beta}} \quad (5)$$

Single Predictor Raw Score Parameter Model: $Y_i = \alpha + \beta X_i + \varepsilon_i$

Single Predictor Raw Score Statistical Model: $\hat{Y} = a + b_1 X_1$

Estimate of Beta (b): $b = \beta \frac{s_Y}{s_X}$ (6)

Since Beta-hat and r are identical in the single predictor model r can be substituted.

Estimate of the Y-intercept or Regression constant (a): $a = \bar{Y} - b\bar{X}$ (7)

Standard error of estimate: $s_{Y.X} = s_Y \sqrt{1 - r_{YX}^2}$ (8)

Standard error of b: $Se_b = \sqrt{\frac{s_{Y.X}^2}{(N-1)s_X^2}}$ (9)

There are two identical null hypotheses: $H_0: \beta = 0$ and $H_0: \rho = 0$

Both are tested with a t -statistic with $(N - 2)$ degrees of freedom (df) which can be computed two ways. Again with formula (4) and with

$$t_{(N-2)} = \frac{b - 0}{Se_b} \quad (10)$$

Two Predictor Standardized Parameter Model: $Z_{Yi} = \beta_1 Z_{X1i} + \beta_2 Z_{X2i} + \varepsilon_i$

Two Predictor Standardized Statistical Model: $\hat{Z}_{Yi} = \hat{\beta}_1 Z_{X1i} + \hat{\beta}_2 Z_{X2i}$

To calculate Beta-hat the correlation between the predictor variables must be taken into consideration

$$\hat{\beta}_1 = \frac{r_{Y1} - r_{Y2} r_{12}}{1 - r_{12}^2} \quad (11) \quad \text{and} \quad \hat{\beta}_2 = \frac{r_{Y2} - r_{Y1} r_{12}}{1 - r_{12}^2} \quad (12)$$

Similar to formula (2), the Standard error of estimate is:

$$s_{Zy.\hat{Z}_y} = \sqrt{1 - R_{YX}^2} \quad (13)$$

In the two predictor case the Standard error of Beta-hat is the same for both

variables:

$$Se\beta = \sqrt{\frac{(1 - R_{Y.12}^2)}{(N-3)(1 - r_{12}^2)}} \quad (14)$$

However, there is more than one null hypothesis that can be tested.

First of all, one can test whether the overall model significantly improves prediction over the mean. $H_0: \beta_1 = \beta_2 = 0$

This is tested with an F-statistic with two (number of predictors) and (N-3) *dfs*:

$$F_{(2,N-3)} = \frac{(R^2 - 0)/2}{(1 - R^2)/(N - 3)} \quad (15)$$

Multiple R^2 has a general formula:

$$R_Y^2 = \sum_{j=1}^K \hat{\beta}_j r_{Yj} = \hat{\beta}_1 r_{Y1} + \hat{\beta}_2 r_{Y2} + \dots + \hat{\beta}_k r_{Yk} \quad (16)$$

One may also test whether each predictor makes a significant improvement in prediction over the other predictor(s). This is tested with a t-test with (N - k - 1) degrees of freedom, where k equals the number of predictors (in this case k = 2).

For any variable j :

$$t_{(N - k - 1)} = \frac{\hat{\beta}_j - 0}{Se\beta_j} \quad (17)$$

where $Se\beta_j = \sqrt{\frac{(1 - R_{Y.12 \dots k}^2)}{(N-k-1)(1 - R_{j.12 \dots k}^2)}} \quad (18)$

This can also be tested with a more flexible F-statistic:

$$F_{(k_F - k_R, N - k_F - 1)} = \frac{(R_F^2 - R_R^2)/(k_F - k_R)}{(1 - R_F^2)/(N - k_F - 1)} \quad (19)$$

Two Predictor Raw Score Parameter Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

Two Predictor Raw Score Statistical Model: $\hat{Y} = a + b_1 X_1 + b_2 X_2$

For any variable j , the Estimate of Beta (b_j): $b_j = \hat{\beta}_j \frac{s_Y}{s_{X_j}}$

(20)

Estimate of the Y-intercept or Regression constant (a): $a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$ (21)

Similar to formula (8), the Standard error of estimate: $s_{Y.\hat{Y}} = s_Y \sqrt{1 - R^2}$ (22)

Because of possible differences in variance across variables, each predictor variable has a different Standard error of b:

For any of the two variables denoted as j : $Se_{b_j} = \sqrt{\frac{s_{Y.\hat{Y}}^2}{s_j^2(N-1)(1 - r_{12}^2)}}$ (23)

Again, one can test whether the overall model significantly improves prediction over the mean. $H_0: \beta_1 = \beta_2 = 0$, which is tested with the F -statistic in formula (15).

Also similar to the standardized model, one may also test whether each predictor makes a significant improvement in prediction over the other predictor(s). This is tested with a t-test with $(N - k - 1)$ degrees of freedom, where k equals the number of predictors (in this case $k = 2$).

For any variable j : $t_{(N - k - 1)} = \frac{b_j - 0}{Se_{b_j}}$ (24)

Partial Correlations are used to statistically "control" the effects of all other predictors. Partial correlations remove the effect of control variables from variables of interest including the dependent variable. Some researchers use them instead of Beta-hat to interpret variable "importance."

With one dependent variable (Y) and two predictors, the general formula is:

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2} r_{12}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{Y2}^2}} \quad (25)$$

Semi-Partial (sometimes referred to as Part) correlation are an index of the "unique" correlation between variables. Semi-Partial correlations remove the effect of a variable from all other predictors but not the dependent variable.

With one dependent variable (Y) and 2 predictors, the general formula is:

$$r_{Y(1.2)} = \frac{r_{Y1} - r_{Y2} r_{12}}{\sqrt{1 - r_{12}^2}} \quad (26)$$

Squaring Semi-partial correlations are useful because they give the "unique" contribution a variable makes to the R^2 of a multiple regression model.

For example with two predictors R^2 can be decomposed as follows:

$$R_{Y.12}^2 = r_{Y2}^2 + r_{Y(1.2)}^2 \quad \text{and conversely,} \quad R_{Y.12}^2 = r_{Y1}^2 + r_{Y(2.1)}^2$$

Source Table for Multiple Regression

Although this process would be laborious, this is the conceptual derivation for the F -ratio in Multiple Regression

Source	Sum of Squares	df	Mean Squares	F
Regression (Explained Variance)	$\sum (\hat{Y}_i - \bar{Y})^2$	k	SS_R/k	MS_R/MS_e
Residual (Error Variance)	$\sum (Y_i - \hat{Y}_i)^2$	$N - k - 1$	SS_e/df_e	
Total Variance	$\sum (Y_i - \bar{Y})^2$	$N - 1$	$s^2 = SS_T/N-1$	

where, N = total number of cases, k = number of predictors, \bar{Y} = the mean of Y .
 Y_i = each individual score on Y , and \hat{Y}_i = each individual predicted Y .

Given, $R^2 = SS_R/SS_T$

Source	Sum of Squares	df	MS	F
Regression (Explained Variance)	$R^2 SS_T$	k	SS_R/k	$\frac{(R^2/k)}{(1 - R^2)/(N - k - 1)}$
Residual (Error Variance)	$(1 - R^2) SS_T$	$N - k - 1$	$SS_e/N - k - 1$	
Total Variance	$\sum (Y_i - \bar{Y})^2$	$N - 1$	$s^2 = SS_T/N-1$	

One-Way ANOVA Source Table

When we extend Least Squares Regression Methodology to a continuous dependent variable Y and categorical independent variables, it is often referred to as the ANalysis Of Varaince (ANOVA). In the ANOVA, the predicted score, \hat{Y}_i , for each individual in the j th group is equal to their (j)th group mean, $\hat{Y}_i = \bar{Y}_j$. Knowing this, the previous Source Tables simplify greatly. For the One-way (one categorical independent variable) ANOVA, the Source Table is as follows:

Source	Sum of Squares	df	Mean Squares	F
Between Groups (Explained Variance)	$\sum n_j(\bar{Y}_j - \bar{Y}^*)^2$	$J - 1$	$SS_B/J - 1$	MS_B/MS_W
Within Groups (Error Variance)	$\sum (Y_i - \bar{Y}_j)^2$	$N - J$	SS_W/df_W	
Total Variance	$\sum (Y_i - \bar{Y}^*)^2$	$N - 1$	$s^2 = SS_T/N-1$	

where, N = total number of cases, J = number of groups, \bar{Y}^* = the grand mean of Y across all groups. Y_i = each individual score on Y , and \bar{Y}_j = the mean for group j .
 n_j = the number of cases in group j .

$R^2 = \eta^2 = SS_B/SS_T$.