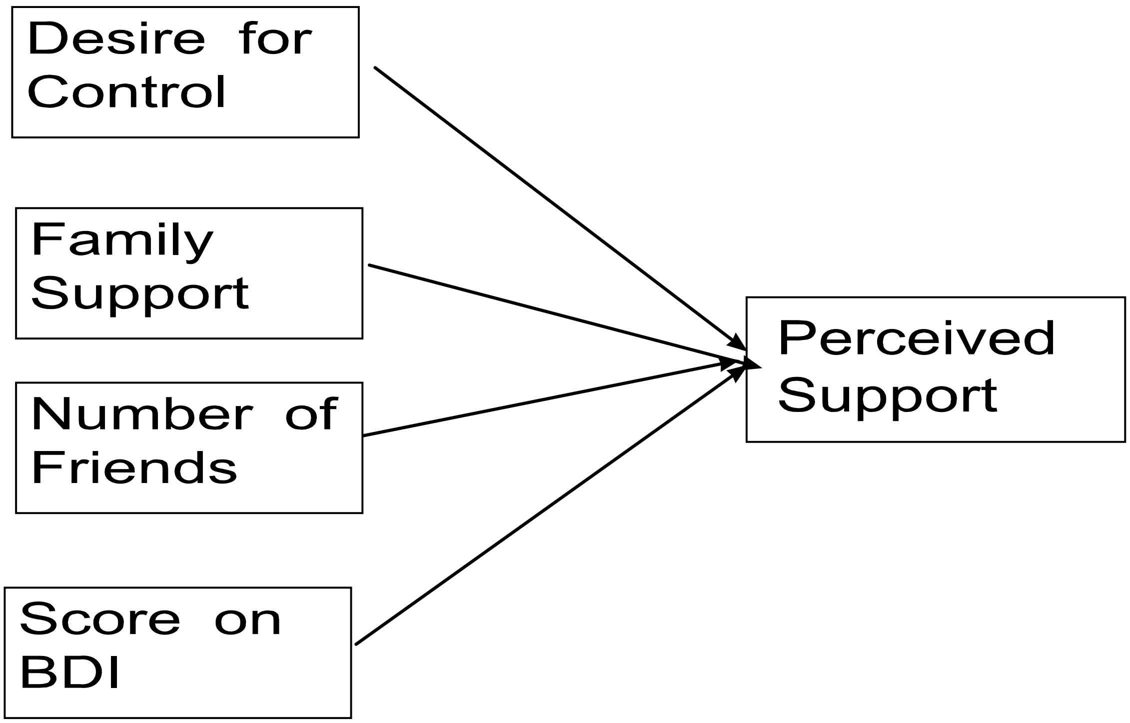
Multiple Regression

Problem: we want to determine the effect of Desire for control, Family support, Number of friends, and Score on the BDI test on **Perceived Support** of Latino women.



- **Dependent variable**: Perceived support.
- **Independent Variable** 1: Desire for control. Measured through a questionnaire.
- **Independent Variable** 2: Family support. Measured through a questionnaire.
- **Independent Variable** 3: Number of friends.
- **Independent Variable** 4: Score on the BDI.

Questions that we may have about the variables

- Is the relationship between Perceived support (**DV)** and Desire for control **(IV**1) the same when we use a simple model than when we also include: *Family support* **(IV**2)? How about when we include *Number of friends* **(IV**3)? etc.

- **It depends on how correlated the variables are. For most conditions, it is not.**

- We need to translate our causal relationship into a mathematical model.

  - Develop an equation:

$$Y = b0 + b1X1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + e$$
$$Y = BX + E$$

or:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} \times \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \\ 1 & x_{31} & x_{32} & x_{33} & x_{34} \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_n \end{pmatrix}$$

Where ***Y*** is an **(n × 1)** vector with the measures for the dependent variable; ***B*** is an **(5 × 1)** vector that contains the coefficients, ***X*** is an **(n × 5)** matrix that contains the measures of the independent variables (the extra column is a vector of "ones" so we can calculate the intercept), and finally, ***E*** is a **(n × 1)** vector that contains the error terms:

- We have to find the parameters of the model (i.e., solve the unknowns in the model, or more formally, compute the solution).

  - we can “solve for b” in our equation:

$$XB = Y$$
$$X'XB = X'Y$$
$$(X'X)^{-1}(X'X)B = (X'X)^{-1}X'Y$$
$$IB = (X'X)^{-1}X'Y$$

- What is the real meaning of the values we get from solving for B?

  - The model is taking into account the level of redundancy among variables as it calculates the best estimates. Therefore, some books call them “**partial regression coefficients**”: the slopes are calculated to include the influence of other variables in the model. For example, for our model with four IVs, the first three coefficients can be interpreted as:

- **b_0** estimates the mean of **Y** when **X_1, X_2, X_3 and, X_4 are zero**. This only makes sense when the ranges of both X_1 to X_4 can include 0.

- **b_1** estimates the expected change in ***Y*** when we hold **X_2, X_3 and X_4, constant**.

- Similarly, **b_2** explains the expected change in ***Y*** when we hold **X_1, X_3, and X_4, constant**.

- How good is the fit of the model:

  - Estimation of the *residuals* (difference between ***observed*** and ***predicted*** scores) and the Residual Sum of Squares (RSS):

$$e_i = (y_i - \hat{y}_i)$$
$$RSS = \sum (y_i - \hat{y}_i)^2$$
$$RSS = \sum (e_i)^2$$

  - Estimation of R^2 (**percentage of variance explained**):

$$R^2 = \frac{sd(\hat{y})}{sd(y)} = \frac{\sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{n-1}}}{\sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

  - *Adjusted R^2 value:* **takes the extra regressors into account**:

$$Adj\_R^2 = R^2 - \left(\frac{k-1}{n-k}\right)(1 - R^2)$$

Where k = number of "b's " in the model. Adjusted R^2 gives an estimate of the real change in amount of variance explained due to adding a new regressor to the model. We can also say that adjusted R^2 evaluates if the improvement in the model is small relative to the increase in complexity.

  - *Multiple correlation coefficient* (**measures the association between the DV and an optimal combination of the IV's**):

$$mult\_corr = r_{\hat{Y},Y}$$

- Test of significance: *Omnibus test* (**checks if at least one of the slopes is significant**):

$$F = \frac{MS_{REGRESSION}}{MS_{ERROR}} = \frac{\dfrac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{p-1}}{\dfrac{\sum\left(y_i - \hat{Y}\right)^2}{n-p}}$$ with (k-1, n-k) degrees of freedom.

- The *standard error of the coefficients* is one of the by-products of the matrix approach:

$$se\_B_i = \left(sdev_{resid}\right)^2 \left(X'X\right)_{ii}^{-1}$$

Where $\left(X'X\right)_{i,i}^{-1}$ represents the corresponding element of the main diagonal of the inverse matrix of crossproducts, and $(sdev_{(resid)})^2$ is the standard deviation of the residuals raised to the square (same as the Mean of Squares of Error).

- The *t-test*:

$$t\_b_i = \frac{b_i}{se\_B_i}$$ with (n-k) degrees of freedom.

- *Confidence Intervals* for all the b's:

$$CI\_b_i(1-\alpha) = b_i \pm (t_{tables})\,(se\_B_i)$$

- Confidence Intervals mean of a predicted value **(answers the question: What is the Confidence Interval for the mean (Y) when (X_1, X_2, X_3, X_4) are…)**:

- The tricky part is figuring out the standard error, because we have several IV's. Ask me (or check: Montgomery, D. C., & Peck, E. A. (1982). Introduction to linear regression analysis. NY: John Wiley., pages 127-128).

$$CI\_\hat{Y}_i(1-\alpha) = \hat{Y}_i \pm (t_{tables})\,(se\_\hat{Y}_i)$$

- What if the Independent Variables are correlated?

- ***MULTICOLLINEARITY***: any or all of the IVs are linearly related with any or all of the others. Sources of Multicollinearity:

        - **Data collection method**: when we sample only a limited region of the population. By doing so, we may end up with **strongly-correlated** variables.

        - **Constraints on the model**: In this case, it does not matter how we sample, we will always get that constraint

        - **Choice of the model**: models that use polynomial terms (like age^2), in addition to the linear term (i.e., age).

        - **Over defined model**: a model with more IV than cases. Very common in psychology and health sciences (e.g., clinical cases).

- What if we have multicollinearity?

    - If we have multicollinearity, we may have a misleading interpretation of the regression coefficients (coefficients cannot be trusted).

        - The principal problem with this estimates is the **extrapolation to other samples/other values beyond those used to estimate the coefficients**. The coefficients are unreliable because they will change from sample to sample.

        - If we have multicollinearity, t**he standard error of the coefficients will be huge**. Thus, slight <u>different samples will give very different estimates of the same coefficient.</u>

        - Theoretically (i.e., after an infinite number of samples are taken), the value of the coefficients will converge to the mean. However, **in any given sample, the value may be way off… Even, of opposite sign!**

        - Because of the huge standard error, and inaccurate estimation of the coefficients, **we loose power** (i.e., it is harder to reject the null hypothesis that the b_i's are different from zero).

- How can we spot Multicollinearity?

    - **Check the correlation matrix.** If we find large correlations among independent variables, then we know that we have the problem.

- Check the **determinant of the (X'X) matrix**.

- **The values of the main diagonals of the Inverse of the Correlation matrix among IV's $(\mathbf{C}^{\mathbf{T}}\mathbf{C})^{-1}$** matrix are equal to:

$$\left(C^T C\right)^{-1}_{i,i} = \frac{1}{1 - R^2_{(bi)}}$$

Where $R^2_{(bi)}$ is the coefficient of determination we get when X_i was regressed on the remaining p-1 regressors. The elements of the main diagonal are the so-called **Variance Inflation Factor (VIF)** (reported by SPSS). The value:

$$1 - R^2_{(bi)}$$

Is called **tolerance.** We can see that we want this value to be close to 1**,** because that means that $R^2_{(bi)}$ *is almost zero***.** This value is also reported in SPSS.

- **Check the value of the standard error of the coefficients** (compare it to se$_{(bi)}$ when it is the only independent variable in the model)**.**

- **Compare the significance values of both the F and the t's**. Multicollinearity sometimes makes the **F-test to be significant**, while the t's are not (because the standard errors of the coefficients are huge).

- **The signs and magnitudes of the regression coefficients can also sometimes provide an indication of multicollinearity**. If adding or removing an IV produces wild changes in the estimates, then there is multicollinearity.

- In addition, if **deletion of one or more data points produces wild variations in the coefficients**, that may be an indication of multicollinearity.

- If the values of the **standardized regression coefficients** are larger than either +1 or -1, it means that we have problems of multicollinearity.

- If the **signs of the coefficients are contrary to what you know/expect**, then be alert about the possibility of multicollinearity.

- If the **condition index** for one of the eigenvalues is too high, then we may have a problem with multicollinearity.
- If a high proportion of the variance of two or more coefficients is associated with the same **eigenvalue**, <u>then that is a clear index of multicollinearity.</u>

- What to do if we have multicollinearity?
  - **Get rid of some of the variables** that are creating the problems.
  - Try to **combine their scores into one single value**
  - Keep the model, under the understanding that **generalizations beyond the sample are risky.**
    - **Do not try to interpret the b's.**
    - OK as long as multicollinearity is an integral part of the model (the population always shows the same level of relation among independent variables).
    - CI for prediction are not affected. **However, do not try to predict outside the ranges of your variables in the model.** Prediction is better if close to the means of the variables.
  - **If polynomial models**, try centering them, or use **orthogonal polynomials**.
  - Use **hierarchical models**.
  - Try to find-out what is the latent construct behind the correlated variables (do **factor analysis**).
  - Try the technique called **Ridge Regression**, which is more robust to multicollinearity.
  - **Add more cases**. This of course in case you suspect that the multicollinearity problem is due to sampling bias (check causes for multicollinearity).