

HARRY GANZEBOOM'S GUIDE TO READING SPSS REGRESSION TABLES

This version: February 5 2009. Revised: October 28 2013.

Syntax

```
Regr /dep=Y /enter=X1 X2 /enter=X3.
```

This spss statement generates a stepwise series of regression models, in which you can enter ever more X-variables. You determine the (forward) steps by the **/enter** statements. The output is repeated for each model separately.

By default, there is listwise deletion of missing values and your results are evaluated by the largest possible N among the X-vars (which may be very small). You can use **/missing=pairwise** to learn more about occurrence and influence of missing values.

Other useful options and statistics:

```
/des=uni corr n  
/stat=tol collin change default.
```

Coefficients

- Start at the lower left corner of the Unstandardized Coefficients B. This is the regression model itself.
- First interpret the intercept (Constant): this is the expected value of the dependent variable Y for cases that are 0 on each of the independent variables X. When this ($X_1=X_2=\dots=0$) never happens in your data, the intercept is an extrapolation and often not of substantive interest. If you want to make sure this does not happen, recode all your X-variables in such a way that 0 occurs as a value I each of them.
- Then interpret the other B's. They show the expected change of Y, when the respective X changes 1 unit.
 - The interpretation of B's is partial: the expected change of Y for 1 unit of X holds when all the other X-variables in the model are (statistically) controlled (=kept constant).
 - B's can only be interpreted relative to one another if the units of the respected X-variables can be compared to one another. Example: B's of age and sex are not comparable, B's of father's education and mother's education may be comparable.
 - If you do not like the scale of numbers that you see, because they have too many or too few decimals, consider changing the unit of measurement, so that 1 unit becomes interpretable. For example, consider changing the scale of X to a 0..1 range: the B then expresses the change of Y over the entire range of X. This does not change the contents of the model.
- Move to the column "Standardized Coefficients" with the beta's. Like the B's, the beta's tell you how much change of Y occurs for 1 unit of X, but this time the variables (both X's and Y) are Z-standardized. So, a beta tells you how many units standard deviation of Y you get for 1 standard deviation of X

- Beta's are in the metric of correlations: between -1 and $+1$, with 0 as a neutral point (=no effect).
- In simple regression, the beta is equal to the correlation between X and Y .
- Beta's can be compared to one another, they allow conclusions like: "age is more important than sex in determining Y ".
- Standardized equations have an implicit intercept, 0 . When all the X -variables are 0 (i.e. for a case with mean values on all the X -variables), the expected value of Y would be its mean, being 0 for a standardized Y -variable.
- Standardized effects have a similar purpose as 'effect sizes' in experimental designs. They tell you how strong effects are on 'absolute' scale and how they compare in strength to one another.
- Standard errors (SE) of the B's: Standard errors (or 'sampling errors') give an estimate of how the B's would vary if we were to draw many samples of size N . The SE's are the standard deviations of the sampling distributions of the B's.
 - SE's are in the unit of X : if you change the unit of X , the SE's change in the same way.
 - Using SE's you can construct confidence intervals.
 - SE's become larger (as do confidence intervals)
 - When N is smaller
 - When there are more X -variables in the model; however, they may become smaller if these X -variables are strong predictors of Y .
 - When these X -variables are stronger correlated.
 - When X -variables are extremely correlated (and there is low N), SE's can become very large. This condition is referred to as (multi)collinearity.
- t -values and Sig.: $t = B/SE$. This value should be compared to the t -distribution that you find in any statistics book to find the associated probability. Fortunately, you do not have to do it all the time, because SPSS does this for you. It prints the P that is associated with the t and calls it Sig.
 - Even if it says Sig. = $.000$, the value is not exactly zero, it is just very small.
 - The Sig. column would be more appropriately labeled P .
 - P is the probability that a t -value of this size or more extreme would be observed in a (simple random) sample of size N if the H_0 is true in the population.
 - P -values are two-tailed: if you have a strong expectation about the location of H_a , you can use a one-tailed probability: divide P by 2 .
 - We conclude that a B is statistically significant at the $.05$ significance level, if $P < .05$ (and $t > 1.96$).
 - The choice of the significance level is yours, and customarily made at $.05$ ($1/20$).
 - If you choose a $.05$ significance level, it means that you decide to reject the null hypothesis incorrectly in 5% of all tests you do. This is called the Type I error.
 - Note that SPSS always assumes $H_0=0$. This is often a sensible choice for the B's, but often not for the intercept.

ANOVA table

- The Sum of Squares SS have the following interpretation
 - SS total is the variation of the Y's around their mean.
 - SS residual is the variation of Y's around the regression line/plane.
 - SS regression is the sum of squared deviations between the regression plane (expected values) and the mean Y.
 - $SS\ total = SS\ regression + SS\ residual$
 - If you add more X-variables to the model, SS-regression will increase, SS residual will decrease.
- The degrees of freedom have the following interpretation:
 - $DF\ total = N - 1$
 - $DF\ regression = \text{number of X-variables}$
 - $DF\ residual: DF\ total - DF\ regression$
- Mean squares: SS / DF .
 - Note there is no MS total printed. It would be the same as the variance of Y.
 - $F = MS\ regression / MS\ residual$. This F-value tests whether there is any predictive power in the X-variables as a whole. The F-test is somewhat superfluous because it is almost always statistically significant, and it is always significant when at least some of the t-tests for the B-coefficients are statistically significant.
 - However, F-tests can be useful when there is multi-collinearity, or when comparing different models.
 - The F-test refers to F-distributions as found in most statistics books, with DF regression and DF residual as coordinates. Fortunately, you do not have to do this all the time, because SPSS prints the associated P-value.
 - In simple regression (one predictor), the F-test is equivalent to the t-test for the B. In this case ($DF\ regression = 1$), the F is the square of the t, and has a critical value of $3.84 = 1.96^2$.

Model summary

- The R-square gives the proportion explained variance: $R^2 = SS\ regression / SS\ total$. This is the most commonly used measure of fit of the model.
 - The adjusted R-square corrects the R-square for the fact that adding any (even nonsensical) X-variables to the model would increase the SS regression. If you add a nonsensical variable to the model, the expected increase of SS regression is MS residual. Adjusted R square is therefore defined as: $(SS\ regression - (DF\ reg * MS\ regr)) / SS\ total$.
- In other, related models ('analysis of variance') SS regression is often referred to as SS model. SS residual is often also called the SS error.
- Finally, the Standard Error of the Estimate:
 - Is not a standard error and not related the sampling distribution.
 - It is the standard deviation of the residuals and also a measure of fit: small values mean that all Y's are close to the regression line / plane.
 - A more appropriate name is: Root Mean Square Residual or the Root Mean Square Error (RMSE).

- Standard Error of the Estimate is thus the root of the Mean Squared Error.

Finally

- All inferential statistics (but not the regression model itself) assume: (A) simple random sampling, (B) normally distributed residuals, (C) homogeneously distributed residuals. If this is not true, do take the numbers for SE, t and P with a grain of salt.
- It really helps to look at descriptives (frequencies) of all X and Y variables before starting interpretation. Not only the mean and standard deviations are important, but also minimum and maximum.

Regression and causal analysis

Regression analysis is used for causal analysis, but is not identical to causal analysis. Regression analysis owes its special role in causal analysis by its capacity to control ('keep constant') multiple variables, with sometimes even a very large number of controls. The causal interpretation depends upon assumptions that have to be satisfied outside regression analysis:

- X can be a cause of Y, not the other way around (no reverse causation)
- X-variables should be distinguished in confounders and intervening variables (=mediators). To examine a causal effect of X1 of Y, you need to control all confounding other X-variables, but not the mediators. Only if your aim is to examine mediation (=indirect effects), you need to control mediators.
- Evidence of causality is when an effect of X1 on Y remains after controlling confounders.

It is often useful to do your regression analysis in forward steps, in which variables are entered (or removed) in groups. The most commonly used logic of analysis is:

1. Model with only a bivariate effect of X1.
2. Control confounders
3. Control mediators.

Note that the third step is only useful when you are answering a mediation problem.

Regression analysis and analysis of variance

For experimental designs one often uses analysis of variance (ANOVA), and if there are multiple dependent variables MANOVA. While there are special programs to conduct anova, it is in fact identical with linear regression (and spss's UNIANOVA is also useful to do regression analysis). Experimental designs most often concentrate on two uncorrelated discrete (often dichotomous) X-variables and the question of interest is about the interaction effect (what is the effect of the combination of X1 and X2 on Y?). In a properly designed experiment there are no questions about controls, confounders, mediators or even effect sizes (=standardized effects).

Some more complicated experimental designs (such as pretest/posttest designs, manipulation checks, or quasi-experimental designs) resemble observation data more strongly. Here stepwise regression is often useful.