## Main Instructions

1. Read the description of the affairs dataset in "affairs-data-description.rtf"
2. Import "affairs.csv" into SPSS
3. Add variable labels and value labels and adjust any other meta data properties
4. Examine frequencies and plots of all variables to get a sense of the data
5. The aim of the analysis will be to predict extra-marital affairs. To do this, we will need to modify several variables: (a) create a binary version of affairs (0 for no affairs, 1 for one or more affairs; (b) convert gender into 0=female; 1 = male; (c) convert children into 0=no, 1=yes)
6. Produce a correlation matrix of all variables (use transformed rather than original variables) and interpret the overall pattern of results
7. Run a binary logistic regression predicting whether someone reports having had an affair in the last year from the other predictors. Interpret the results: overall model prediction, individual coefficients, statistical significance, etc.
8. Provide an overall interpretation of the results.

# Answers

## 1. Read the description of the affairs dataset in "affairs-data-description.rtf"

Get a sense of the variables included and the scales of measurement that they are on.

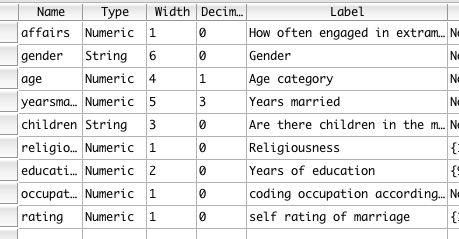
## 2. Import "affairs.csv" into SPSS

The csv file extension stands for comma-separated-values. Use the SPSS Import text file wizard ("File - read text data": "yes variable names are included in top of the file". Save the file as a ".sav" SPSS file.

## 3. Add variable labels and value labels and adjust any other meta data properties

Go into "Variable View" to modify meta-data.

Variable labels come from the data description file. For example, you might add variable labels like the following:



And value labels to any variable that has numbers which represent something else. Note that there is often a choice about whether to represent categorical variables in SPSS as numbers with value labels or as a "string" variable. In particular, when importing from a text data file, string variables are often used.

## 4. Examine frequencies and plots of all variables to get a sense of the data

Run "Analyze - Descriptives - Frequencies " and put all variables in (you may want to add "charts - histograms").

The best choice of univariate plots is often difficult to determine until you know something about the chracteristics of the variable. That said, for categorical variables, frequencies are often sufficient. It's typically numeric variables where plots are most relevant in providing a visual snapshot of distributional properties and the presence of outliers.

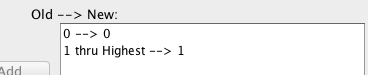
At this stage, you should be making various observations about the data. For example:

* extra marital sex: The majority of the sample reports no extra-marital sex. The large number of people reporting 12 and 7 rather 6,8,9,10,11, etc. suggests that there must have been something in the data collection procedure to lead to this.
* Age: age is also somewhat discretised based on age categories. However, the age distribution looks roughly representative of a typical married adult population, albeit a little younger.
* rating: Most people are happy or very happy with their marriage. About 13% are somewhat or very unhappy.
* etc.

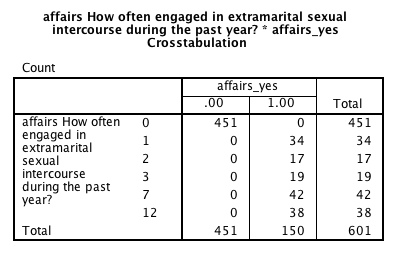
The point of this process is to get a feel for the variables. From this analysis we can characterise this sample. Is it likely to be similar to a general population of married people? What kind of features does it have?

## 5. The aim of the analysis will be to predict extra-marital affairs. To do this, we will need to modify several variables: (a) create a binary version of affairs (0 for no affairs, 1 for one or more affairs; (b) convert gender into 0=female; 1 = male; (c) convert children into 0=no, 1=yes)

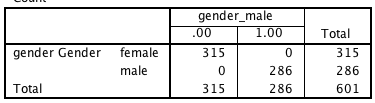
(a) There are several ways to convert a numeric variable into a binary variable. "Transform - recode into different variable: affairs into output variable="affairs\_yes", and click "change"; click "old and new values" and the old value = 0, new value = 0, add; then Range, value through highest = 1, new value = 1, add; continue.



It's also good to get in the habit of checking that your transformations have been successful. E.g., you can create a table (analyze - descriptives - crosstabs put the original variable in rows and the transformed variable in columns):



(b) Use "Transform - recode into different variable": "gender --> gender\_male" (note this naming convention makes it clear that 0 is female and 1 is male). Click old and new values: old: male, new:1; old:female, new=0. (note that this process is case sensitive, so you need to check whether the original variable recorded gender as "Female" or "FEMALE" or "female". Once again do a cross tabs to check.



(c) Repeat as per previous step, the final syntax might look like this.

RECODE children ('yes'=1) ('no'=0) INTO children\_yes.

EXECUTE.

Note that with all these data transformations that you should be pasting them into a syntax file and then run them. Furthermore, as you get more familiar with the syntax, it is typically quicker to manipulate the syntax than to use the menus. However, the main benefit of syntax is reproduciblity.

## 6. Produce a correlation matrix of all variables (use transformed rather than original variables) and interpret the overall pattern of results

Run "analyze - correlate - bivariate"

You may wish to use pivot trays and so on to get just the correlations.

The main aim is get a feel for the pattern of results.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | age Age category | 1.00 | .78 | .19 | .14 | .17 | -.20 | .06 | .19 | .42 |
| 2 | yearsmarried Years married | .78 | 1.00 | .22 | .04 | .05 | -.24 | .14 | .03 | .57 |
| 3 | religiousness Religiousness | .19 | .22 | 1.00 | -.04 | -.04 | .02 | -.13 | .01 | .13 |
| 4 | education Years of education | .14 | .04 | -.04 | 1.00 | .53 | .11 | .02 | .40 | -.01 |
| 5 | occupation coding occupation according to Hollingshead classification (reverse numbering). | .17 | .05 | -.04 | .53 | 1.00 | .02 | .04 | .47 | -.09 |
| 6 | rating self rating of marriage | -.20 | -.24 | .02 | .11 | .02 | 1.00 | -.25 | -.01 | -.20 |
| 7 | affairs\_yes | .06 | .14 | -.13 | .02 | .04 | -.25 | 1.00 | .05 | .13 |
| 8 | gender\_male | .19 | .03 | .01 | .40 | .47 | -.01 | .05 | 1.00 | .07 |
| 9 | children\_yes | .42 | .57 | .13 | -.01 | -.09 | -.20 | .13 | .07 | 1.00 |

A few observations can be made

* The strongest correlation with affairs is low marital satisfaction. However, longer term marriages, being less religious, and the presence of children are all related to affairs. That said, all the correlations are pretty small.
* The correlations between most of the predictors are pretty small. Age and years marriage are strongly related. And the presence of children is related to age and years of marriage. Education and occupational level are also related.
* It's also interesting to inspect what variables relate most to marital satisfaction. There are a number of correlations in the .20 range. In particular, younger people, newer marriages, marriages without affairs, and not having children are associated with greater marriage satisfaction.

In general, you should take from this process several points. Bivariate correlations provide a rich picture of a dataset. There is a story to be learnt about how things interrelate. The emphasis should be on the pattern. What are the strong relationships? What are the weak relationships? Furthermore, when we move to predicting the presence of affairs its good to be mindful of correlations between predictors. For example, it will be difficult to tease out age from years of marriage when interpreting the meaning of regression coefficients.

## 7. Run a binary logistic regression predicting whether someone reports having had an affair in the last year from the other predictors. Interpret the results: overall model prediction, individual coefficients, statistical significance, etc.

For another example of output interpretation see: <http://www.ats.ucla.edu/stat/spss/output/logistic.htm>

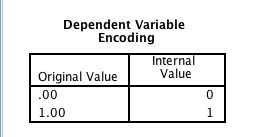
Run "Analyze - regression - binary logistic"

dependent: "affairs\_yes"

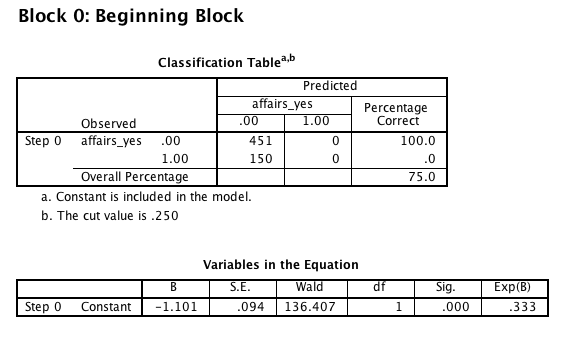
Covariates: age, yearsmarried, religiousneess, education, occupation, rating, gender\_male, children\_yes

Options: Classification plots, and CI for exp(B)

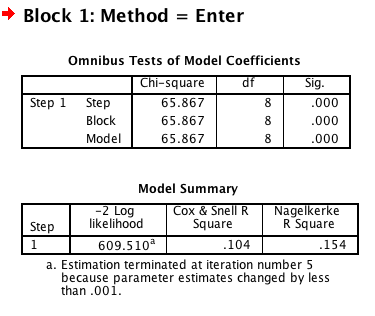
You could also alter the classification cutoff if you wished. By default it is set to 0.5, but you might want to match the sample proportion who have affairs (i.e., around 0.25; thus the cut-off would be 1 - 0.25 = 0.75). I'll do that for these analyses:

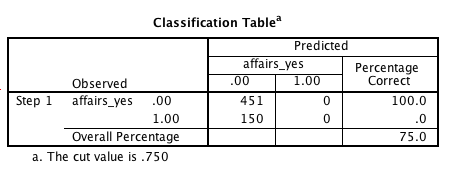


This table confirms that logistic regression has coded the dependent variable as we would expect (i.e., 1 is having an affair; 0 is not having an affair; this is important for understanding what the sign of regression coefficients mean).



Block zero output can mostly be ignored. It is a reference model where the only coefficient is the constant. Thus, it simply predicts that no one had an affair because that was the most common category.

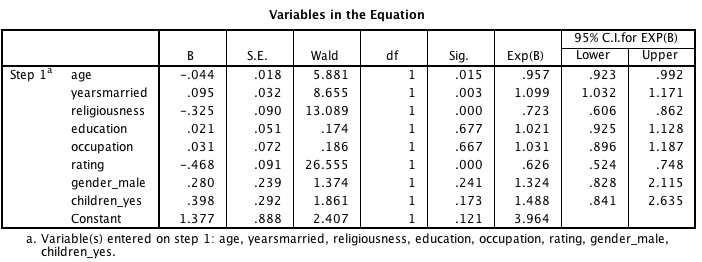




Block 1 is the model we are principally examining.

The omnibus test of model coefficients tests the null hypothesis that the coefficients for all predictors besides the constant are zero. This is rejected in this case. It is analogous to the ANOVA F-test for r-square in multiple regression.

There are several ways to evaluate the overall performance of the model. There are some pseudo-r-square values, however, their interpretation is not always clear. In general lower deviance values (i..e., -2 log likelihood) are desirable. However, this value is only really useful when comparing multiple models. The classification performance is useful. In this case the model was unable to predict better than the null model where everyone is assigned to the not having affairs category. In many respects this is not surprising given that we are predicting an uncommon category and none of the predictors are especially strong at the bivariate level.



In terms of evaluating the predictors, we have the coefficients table.

We can either focus on the B or the Exp(B) columns.

* B refers to the predicting increase in the log odds for a one unit increase in the predictor holding all other predictors constant. Thus, zero means no effect, positive numbers mean more of the predictor mean more likely to be in the focal category, and negative numbers mean less likely to be in the focal category.
* Exp(B) refers to the odds ratio of a one unit increase on the predictor to not increasing the predictor by one unit, holding all other predictors constant. Thus, one means no effect, numbers greater than one mean more of the predictor mean more likely to be in the focal category, and numbers between zero and one mean less likely to be in the focal category.

There are four significant effects. Specifically, people who are more likely to have an affair are: younger (age), married for longer (years married), less religious (religousness), and less satisfied with their marriage (rating).

In terms of specific interpretations

* Moving one unit higher on the 1 to 5 scale of marital satisfaction, leads to a 0.468 reduction in the log odds of having an affair holding all other predictors constant. To give you a sense of the log odds scale, see this table:

|  |  |
| --- | --- |
| ln(p/(1-p) | p |
| -2.20 | 0.1 |
| -1.39 | 0.2 |
| -0.85 | 0.3 |
| -0.41 | 0.4 |
| 0.00 | 0.5 |
| 0.41 | 0.6 |
| 0.85 | 0.7 |
| 1.39 | 0.8 |
| 2.20 | 0.9 |

So for example, a .468 reduction might represent a movement in probabilities of having an affair from 0.5 to around 0.4 or from 0.7 to 0.6.

In terms of identifying relative variable importance, the relative size of the Wald statistic gives some indication. That said, follow-up analyses could be performed where for example, numeric predictors were coded as z-scores in order to put them on a common scale.

## 8. Provide an overall interpretation of the results.

Overall, around a quarter of participants reported having an extra-marital affair. The results suggest that self-report marital affairs are fairly difficult to predict from the variables provided. That said, several variables provided modest prediction. Interestingly, age related to probability of affairs at the bivariate level, but after controlling for years of marriage, the sign of that variable flipped. This suggests that years of marriage is actually the factor driving the effect.

It is not surprising that marital satisfaction is negatively related to having affairs. Presumably this could be a bidirectional relationship. That said, it might be interesting to do follow-up analyses excluding it from the logistic regression or even predicting marital satisfaction.

It is interesting to note that religiousness was one of the stronger predictors of not having an affair. Furthermore, religiousness was pretty much unrelated (or if anything slightly negatively related) to marital satisfaction at the bivariate level.