**Correlation and Regression Analysis: SAS[[1]](#footnote-1)©**

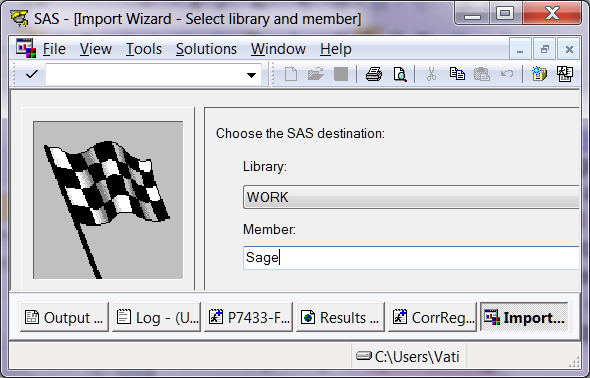
Start this lesson by downloading the required data and program files:

* [Cyberloaf\_Consc\_Age.sav](http://core.ecu.edu/psyc/wuenschk/SPSS/Cyberloaf_Consc_Age.sav)
* [Potthoff.dat](http://core.ecu.edu/psyc/wuenschk/StatData/POTTHOFF.dat)
* [CorrRegr.sas](http://core.ecu.edu/psyc/wuenschk/SAS/CorrRegr.sas)

# Bivariate Analysis: Cyberloafing Predicted from Personality and Age

These days many employees, during work hours, spend time on the Internet doing personal things, things not related to their work. This is called “cyberloafing.” Research at ECU, by Mike Sage, graduate student in [Industrial/Organizational Psychology](http://www.ecu.edu/cs-cas/psyc/IO-Home.cfm), has related the frequency of cyberloafing to personality and age. Personality was measured with a [Big Five](https://www.google.com/search?rlz=1C2GGGE_enUS414US459&num=50&safe=off&site=&source=hp&q=big+five+personality&oq=big+five+personality&gs_l=hp.3..0l10.97045.104685.0.105496.21.12.0.9.9.0.232.1521.1j10j1.12.0.msedr...0...1c.1.64.hp..1.20.1383.0.jmw-YtTw8h4) instrument. Cyberloafing was measured with an instrument designed for this research. Age is in years. The cyberloafing instrument consisted of 23 questions about cyberloafing behaviors, such as “shop online for personal goods,” “send non-work-related e-mail,” and “use Facebook.” For each item, respondents were asked how often they engage in the specified activity during work hours for personal reasons. The response options were “Never,” “Rarely (about once a month),” “Sometimes (at least once a week),” and “Frequently (at least once a day).” Higher scores indicate greater frequency of cyberloafing.

For this exercise, the only Big Five personality factor we shall use is that for Conscientiousness. Bring the data, [Cyberloaf\_Consc\_Age.sav](http://core.ecu.edu/psyc/wuenschk/SPSS/Cyberloaf_Consc_Age.sav), into SAS. See my [brief tutorial](http://core.ecu.edu/psyc/wuenschk/SAS/SPSS2SAS.pdf) on importing SPSS data. Name the member “Sage.” Edit the program properly to point to where you saved the data files and then run the program, [CorrRegr.sas](http://core.ecu.edu/psyc/wuenschk/SAS/CorrRegr.sas).



Look at the output. Before proceeding with the correlation analysis I investigated the variables with an eye to detecting outliers and any violation of the normality assumption. The age variable does show a distinct positive skewness. I could reduce that with a log transformation, but elected not to do so.

**PROC CORR** gives us some simple, univariate (single variable) statistics, and then the correlation coefficients, each with a *p* testing the null hypothesis of no correlation in the population. Recall Cohen’s benchmarks: an *r* of .1 is small, .3 is medium, and .5 is large.

The **PROC REG** statement asks for a regression analysis for each of two models. Proc Reg will continue to run after it has done the requested analysis. This can cause problems if you want to clear the Results window. Invoking the “quit” command prevents this problem.

The first model is for predicting cyberloafing from Conscientiousness. The output gives us the intercept and slope for the least squares regression line, and tells us that the slope is significantly different from zero. Notice that the square of the ***t*** for testing the slope is equal to the ANOVA ***F***. Both test the same null hypothesis. The *R2* indicates that the model explains 31.7 % of the variance in cyberloafing. Sample *R2* tends to overestimate population *ρ2*, so SAS also gives us the ADJusted *R2*, also known as the “**shrunken *R2***,” a relatively unbiased estimator of the population *ρ2*. For a bivariate regression it is computed as:

.

Use my program [Conf-Interval-r.sas](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm) or the [calculator at Vassar](http://vassarstats.net/rho.html) to put a 95 confidence interval on the value of *r* relating cyberloafing to Conscientiousness.

**Presenting the Results**. Subjects were 51 employed graduate students in the School of Business at East Carolina University. Data screening indicated no problems with the assumptions for correlation analysis. Descriptive statistics are shown in Table 1. Cyberloafing was significantly, inversely related to conscientiousness, cyberloafing = 57.04 - .864\*conscientiousness, *t*(49) = 4.77, *p* < .001, *r* = -.563, 95% CI [-.725, -.341].

Table 1  
*Descriptive Statistics for Cyberloafing and Conscientiousness*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | *M* | *SD* | *g1* Skewness | *g2* Kurtosis |
| Cyberloafing | 22.67 | 9.19 | .008 | -.691 |
| Conscientiousness | 39.76 | 5.99 | -.269 | -.882 |

**Trivariate Analysis: Age as a Second Predictor**

I also asked for an analysis based on a second model (model b), with which I predicted cyberloafing from Conscientiousness and from age. Such an analysis, where we have two or more predictor variables, is called a multiple regression.

I have commented out (prevented from being executed, by preceding with an asterisk) the statement which would compute confidence intervals for predicted scores. That would just use up too much screen space with this relatively large data set.

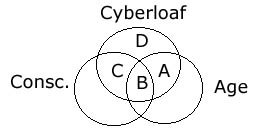
When you look at the output for this **multiple regression**, you see that the two predictor model does do significantly better than chance at predicting cyberloafing, *F*(2, 48) = 20.91, *p* < .001. The ***F*** in the ANOVA table tests the null hypothesis that the multiple correlation coefficient, ***R***, is zero in the population. If that null hypothesis were true, then using the regression equation would be no better than just using the mean for cyberloafing as the predicted cyberloafing score for every person. Clearly we can predict cyberloafing significantly better with the regression equation rather than without it, but do we really need the age variable in the model? Is this model significantly better than the model that had only Conscientiousness as a predictor? To answer that question, we need to look at the "Coefficients," which give us measures of the partial effect of each predictor, above and beyond the effect of the other predictor(s).

The regression equation gives us two unstandardized slopes, both of which are partial statistics. The amount by which cyberloafing changes for each one point increase in Conscientiousness, above and beyond any change associated with age, is -.779, and the amount by which cyberloafing changes for each one point increase in age, above and beyond any change associated with Conscientiousness, is -.276. The intercept, 64.07, is just a reference point, the predicted cyberloafing score for a person whose Conscientiousness and age are both zero (which are not even possible values). The "Standardized Coefficients" (usually called beta, ***β***) are the slopes in standardized units -- that is, how many standard deviations does cyberloafing change for each one standard deviation increase in the predictor, above and beyond the effect of the other predictor(s).

The regression equation represents a plane in three dimensional space (the three dimensions being cyberloafing, Conscientiousness, and age). If we plotted our data in three dimensional space, that plane would minimize the sum of squared deviations between the data and the plane. If we had a 3rd predictor variable, then we would have four dimensions, each perpendicular to each other dimension, and we would be out in **hyperspace**.

The ***t*** testing the null hypothesis that the intercept is zero is of no interest, but those testing the partial slopes are. Conscientiousness does make a significant, unique, contribution towards predicting AR, *t*(48) = 4.759, *p* < .001. Likewise, age also makes a significant, unique, contribution, *t*(48) = 3.653, *p* = .001 Please note that the values for the partial coefficients that you get in a multiple regression are highly dependent on the context provided by the other variables in a model. If you get a small partial coefficient, that could mean that the predictor is not well associated with the dependent variable, or it could be due to the predictor just being highly redundant with one or more of the other variables in the model. Imagine that we were foolish enough to include, as a third predictor in our model, students’ score on the Conscientiousness and age variables added together. Assume that we made just a few minor errors when computing this sum. In this case, each of the predictors would be highly redundant with the other predictors, and all would have partial coefficients close to zero. Why did I specify that we made a few minor errors when computing the sum? Well, if we didn’t, then there would be total redundancy (at least one of the predictor variables being a perfect linear combination of the other predictor variables), which causes the intercorrelation matrix among the predictors to be singular. Singular intercorrelation matrices cannot be inverted, and inversion of that matrix is necessary to complete the multiple regression analysis. In other words, the computer program would just crash. When predictor variables are highly (but not perfectly) correlated with one another, the program may warn you of **multicollinearity**. This problem is associated with a lack of stability of the regression coefficients. In this case, were you randomly to obtain another sample from the same population and repeat the analysis, there is a very good chance that the results (the estimated regression coefficients) would be very different.

Perhaps it will help to see a Venn diagram. Look at the [ballantine](http://www.amstat.org/publications/jse/v10n1/kennedy.html) below.



The top circle represents variance in cyberloafing, the right circle that in age, the left circle that in Conscientiousness. The overlap between circle Age and Cyberloaf, area A + B, represents the ***r2*** between cyberloafing and age. Area B + C represents the *r2* between cyberloafing and Conscientiousness. Area A + B + C + D represents all the variance in cyberloafing, and we standardize it to 1. Area A + B + C represents the variance in cyberloafing explained by our best weighted linear combination of age and Conscientiousness, 46.6% (*R2*). The proportion of all of the variance in cyberloafing which is explained by age but not by Conscientiousness is equal to: .

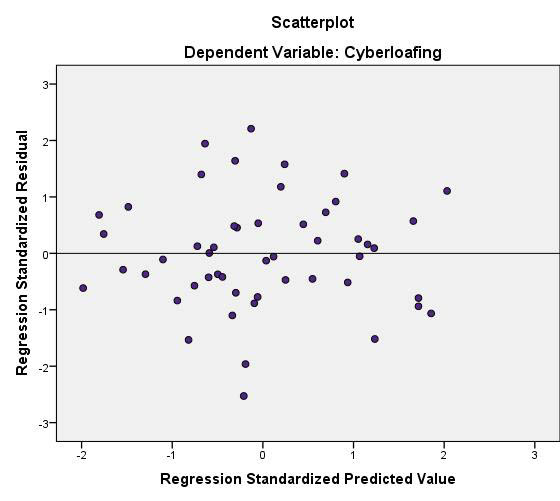
Area A represents the **squared semipartial correlation** for age (.149). Area C represents the squared semipartial correlation for Conscientiousness (.252). Although I generally prefer semipartial correlation coefficients, some persons report the **squared partial correlation coefficients**, which will always be at least as large as the semipartial, and almost always larger. In our Venn diagram, the squared partial correlation coefficient for Conscientiousness is represented by the proportion . That is, of the variance in cyberloafing that is not explained by age, what proportion is explained by Conscientiousness? Or, put another way, if we already had age in our prediction model, by what proportion could we reduce the error variance if we added Conscientiousness to the model? If you consider that (C + D) is between 0 and 1, you should understand why the partial coefficient will be larger than the semipartial.

If we take age back out of the model, the *r2* drops to .317. That drop, .466 - .317 = .149, is the squared semipartial correlation coefficient for age. In other words, we can think of the squared semipartial correlation coefficient as the amount by which the *R2* drops if we delete a predictor from the model.

If we refer back to our Venn diagram, the *R2* is represented by the area A+B+C, and the redundancy between misanthropy and idealism by area B. The redundant area is counted (once) in the multiple *R2*, but not in the partial statistics.

The “Fit Diagnostics” provided by SAS help us see if assumptions have been met and if there are any cases that have unusually large amounts of error of prediction or which have great influence on the solution. I’ll be going into this in much greater detail later.

The plot of the standardized residuals (standardized difference between actual cyberloafing score and that predicted from the model) versus standardized predicted values allows you to evaluate the **normality and homescedasticity assumptions** made when testing the significance of the model and its parameters. If the **normality** assumption has been met, then a vertical column of residuals at any point on that line will be normally distributed. In that case, the density of the plotted symbols will be greatest near that line, and drop quickly away from the line, and will be symmetrically distributed on the two sides (upper versus lower) of the line. If the **homoscedasticity** assumption has been met, then the spread of the dots, in the vertical dimension, will be the same at any one point on that line as it is at any other point on that line. Thus, a residuals plot can be used, by the trained eye, to detect violations of the assumptions of the regression analysis. The trained eye can also detect, from the residual plot, patterns that suggest that the relationship between predictor and criterion is not linear, but rather **curvilinear**.



I opine that you should obtain and report a confidence interval for the value of R2. See [Confidence Intervals for *R* and *R2*](http://core.ecu.edu/psyc/wuenschk/docs30/CI-R2.docx)

# Importance of Looking at a Scatterplot Before You Analyze Your Data

The next section of our program is designed to convince you that it is very important to look at a plot of your data prior to conducting a linear correlation/regression analysis. If you will look at the output, you will see that for each of our four pairs of variables (x1, y1; x2, y2; x3, y3; x4, y4), all of the statistics (mean and *sd* of X, mean and *sd* of Y, the *r*, the slope, the intercept, etc.) are the same as they are for the other three pairs -- but the plots show that the relationship between X and Y is really quite different across these pairs. In the first set, we have a plot that looks about like what we would expect for a moderate to large positive correlation. In the second set we see that the relationship is really curvilinear, and that the data could be fit much better with a curved line (a polynomial function, quadratic, would fit them well). In the third set we see that with the exception of one outlier, the relationship is nearly perfect linear. In the final set we see that the relationship would be zero if we eliminated the one extreme outlier -- with no variance in X, there can be no covariance with Y.

# Moderation Analysis

Sometimes a third variable moderates (alters) the relationship between two (or more) variables of interest. You are about to learn how to conduct a simple moderation analysis.

One day as I sat in the living room, watching the news on TV, there was a story about some demonstration by animal rights activists. I found myself agreeing with them to a greater extent than I normally do. While pondering why I found their position more appealing than usual that evening, I noted that I was also in a rather misanthropic mood that day. That suggested to me that there might be an association between misanthropy and support for animal rights. When evaluating the ethical status of an action that does some harm to a nonhuman animal, I generally do a cost/benefit analysis, weighing the benefit to humankind against the cost of harm done to the nonhuman. When doing such an analysis, if one does not think much of humankind (is misanthropic), e is unlikely to be able to justify harming nonhumans. To the extent that one does not like humans, one will not be likely to think that benefits to humans can justify doing harm to nonhumans. I decided to investigate the relationship between misanthropy and support of animal rights.

Mike Poteat and I developed an animal rights questionnaire, and I developed a few questions designed to measure misanthropy. One of our graduate students, Kevin Jenkins, collected the data we shall analyze here. His respondents were students at ECU. I used reliability and factor analysis to evaluate the scales (I threw a few items out). All of the items were Likert‑type items, on a 5‑point scale. For each scale, we computed each respondent's mean on the items included in that scale. The scale ran from 1 (strongly disagree) to 5 (strongly agree). On the Animal Rights scale (AR), high scores represent support of animal rights positions (such as not eating meat, not wearing leather, not doing research on animals, etc.). On the Misanthropy scale (MISANTH), high scores represent high misanthropy (such as agreeing with the statement that humans are basically wicked).

The idealist is one who believes that morally correct behavior always leads only to desirable consequences; an action that leads to any bad consequences is a morally wrong action. Thus, one would expect the idealist not to engage in cost/benefit analysis of the morality of an action -- any bad consequences cannot be cancelled out by associated good consequences. Thus, there should not be any relationship between misanthropy and attitude about animals in idealists, but there should be such a relationship in nonidealists (who do engage in such cost/benefit analysis, and who may, if they are misanthropic, discount the benefit to humans).

Accordingly, a proper test of my hypothesis would be one that compared the relationship between misanthropy and attitude about animals for idealists versus for nonidealists. Although I did a more sophisticated analysis than is presented here (a "Potthoff analysis," which I cover in my advanced courses), the analysis presented here does address the question I posed. Based on a scores on the measure of idealism, each participant was classified as being either an idealist or not an idealist. Now all we need to do is look at the relationship between misanthropy and idealism separately for idealists and for nonidealists.

The output for the nonidealists shows that the relationship between attitude about animals and misanthropy is significant ( *p* < .001) and of nontrivial magnitude ( *r* = .364, *n* = 91). The plot shows a nice positive slope for the regression line. With nonidealists, misanthropy does produce a discounting of the value of using animals for human benefit, and, accordingly, stronger support for animal rights. On the other hand, with the idealists, who do not do cost/benefit analysis, there is absolutely no relationship between misanthropy and attitude towards animals. The correlation is trivial ( *r* = .020, *n* = 63) and nonsignificant ( *p* = .87), and the plot shows the regression line to be flat.

You can find a paper based on these data at:

<http://core.ecu.edu/psyc/wuenschk/Animals/ABS99-ppr.htm>

# Additional Analyses

The analysis we just completed suggests that nonidealists differ from idealists with respect to the relationship between misanthropy and attitude towards animals, with the *r* and the slope being higher for the nonidealists than for the idealists -- but we have not yet directly tested the null hypotheses that these two groups do not differ on *r* and slope. Please read my document [Comparing Correlation Coefficients, Slopes, and Intercepts](http://core.ecu.edu/psyc/wuenschk/docs30/CompareCorrCoeff.pdf) to see how to test these hypotheses.

Please note that the relationship between X and Y could differ with respect to the slope for predicting Y from X, but not with respect to the Pearson *r*, or vice versa. The Pearson *r* really measures how little scatter there is around the regression line (error in prediction), not how steep the regression line is. [[See relevant plots]](http://core.ecu.edu/psyc/wuenschk/docs01/regr1.gif)

On the left, we can see that the slope is the same for the relationship plotted with blue o’s and that plotted with red x’s, but there is more error in prediction (a smaller Pearson *r* ) with the blue o’s. For the blue data, the effect of extraneous variables on the predicted variable is greater than it is with the red data.

On the right, we can see that the slope is clearly higher with the red x’s than with the blue o’s, but the Pearson *r* is about the same for both sets of data. We can predict equally well in both groups, but the Y variable increases much more rapidly with the X variable in the red group than in the blue.

Additional Reading

* [Annotated Output](http://core.ecu.edu/psyc/wuenschk/SAS/CorrRegrSAS_Output.pdf)
* [Biserial and Polychoric Correlation Coefficients](http://core.ecu.edu/psyc/wuenschk/SAS/Biserial-Polychoric.htm)
* [Comparing Correlation Coefficients, Slopes, and Intercepts](http://core.ecu.edu/psyc/wuenschk/docs30/CompareCorrCoeff.pdf)
* [Confidence Intervals on *R2* or *R*](http://core.ecu.edu/psyc/wuenschk/docs30/CI-R2.docx)
* [Tetrachoric Correlation](http://core.ecu.edu/psyc/wuenschk/SAS/Tetrachoric.docx) -- what it is and how to compute it.
* [Producing and Interpreting Residuals Plots in SAS](http://core.ecu.edu/psyc/wuenschk/SAS/Residual-Plots.docx)

[Return to my SAS Lessons page.](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Lessons.htm)

Copyright 2018, Karl L. Wuensch - All rights reserved.

[Fair Use of This Document](http://core.ecu.edu/psyc/wuenschk/Fair-Use.htm)

1. © Copyright 2018, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)