Soc 63993, Homework #1 Answer Key: Review of Multiple Regression

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised January 22, 2015

The attached table is from an article in Journal for the Scientific Study of Religion, 1990, Vol.29(3):297-314, *Religious Practice: A Human Capital Approach*, written by Laurence Iannaccone. Read the table carefully and answer the following questions.

1. When the dependent variable is CONTRIBUTE, the unstandardized coefficient of INCOME is 9.025. What does this mean?

This means that, IF

- Income increases by one unit, AND
- The values of all the other IVs stays the same, THEN

the average increase in CONTRIBUTE will be 9.025 units. Or, an alternative interpretation would be that, if individuals had identical values on all the IVs except for income, individuals who were one unit higher on INCOME would have an average value on CONTRIBUTE that was 9.025 units higher.

The key idea is that you are looking at the effect of income after "controlling for" other variables, i.e. holding them constant.

Of course, if you did just give somebody $1,000, the values of the other IVs might not stay the same, i.e. increases in INCOME could produce increases or decreases in some other IV which in turn would affect CONTRIBUTE. We'll talk about such relationships more in the Logic of Causal Order.

2. When the dependent variable is CONTRIBUTE, the significance level of INCOME is p≤.001. What does this mean? Which type error does it test? Write out the null and alternative hypotheses that are being tested by the T statistic.

This means that, IF the null hypothesis is true (i.e. $\beta_{Income} = 0$) the odds that a sample would produce an estimated coefficient this large are less than 1 in 1,000. This refers to Type I error, i.e., the probability we falsely reject the null hypothesis when it is true. The null and alternative hypotheses are

H_0: $\beta_{Income} = 0$
H_A: $\beta_{Income} \neq 0$

3. Suppose the researcher believes that older people tend to attend fewer masses than do younger people. Do these results support her?

Obviously not. If the researcher were right, AGE would have a negative effect in the ATTEND equation. The actual estimated effect is positive (0.316).

4. What is the standard error of the coefficient for MARSAME in the CONTRIBUTE equation?

Recall that T = coefficient/standard error, hence standard error = coefficient/T = 70.984/4.60 = 15.43.

5. For the CONTRIBUTE equation, test whether or not R^2 significantly differs from 0 (i.e. compute the F value, its degrees of freedom, and its significance). NOTE: You can use the `Ftail` function in Stata to compute the significance. So, for example, if your computed F = 2.4 with d.f. = 10, 120, the F value is significant at the .0124 level. Ergo, you would reject the null if using the .05 level of significance but not reject if you were using the .01 level.

```
. display Ftail(10, 120, 2.4)
.01236157
```

We are told that N = 555, R^2 = .39, K = 14 (i.e. 14 IVs are listed in the model). Hence,

$$F = \frac{R^2 * (N - K - 1)}{(1 - R^2) * K} = \frac{.39 * (555 - 14 - 1)}{(1 - .39) * 14} = \frac{210.6}{8.54} = 24.66$$

For an F with d.f. = 14, 540, the critical value using even the .01 level of significance is only about 2, so this F value is highly significant. To confirm, using the Stata `Ftail` function [the second `display` command tells Stata to only report the p value to a more easily readable 6 decimal places]

```
. display Ftail(14, 540, 24.66)
2.322e-49
. display %12.6f Ftail(14, 540, 24.66)
0.000000
```

6. When a male and female have identical values on all other IVs, which one will be likely to attend mass more (and by how much)?

In the ATTEND equation, the coefficient for SEX is 7.232. Note that SEX is coded 1 if female, 0 if male. Hence, when males and females have identical values on all other IVs, the female attends an average of 7.232 more masses per year.

7. Suppose you somehow managed to get a score of 0 on all the IVs — how many masses a year would you be expected to attend?

Such people would attend an average of 4.122 masses a year (the value of the constant).

8. Among the three multiple regression equations, three independent variables are not statistically significant in any one of them (HEDUC, NBHD, RAISECA). Why do you think the researchers have still included them in the equations?

It may be that substantive arguments could be made for these variables, and the researchers feared that if they were omitted somebody would argue they should have been in. This way, the researchers explicitly show that the effects of these variables are not significant. Also, note that the researchers are including the same variables in every equation, and that only a few of the regressions they ran are presented. It may be that, in other regressions not shown, these IVs were important.

9. In the ATTEND equation, BELIEF has the largest regression coefficient. Can you therefore say that BELIEF is the most important determinant of Church attendance? Why or why not? What other sorts of information might aid you in determining which variable has the strongest influence?

No. The independent variables are not measured with the same scale, therefore they are not comparable with each other. "Strength of influence" might be better assessed through the standardized coefficients or the partial and semipartial coefficients.

TABLE 1

PARTICIPATION REGRESSIONS:1974 CATHOLIC SURVEY						
	CONTRIBUTE		ATTEND		RATIO (A/C)	
Variable	coefficient	t-stat	coefficient	t-stat	coefficient	t-stat
MARSAME	70.984***	4.60	11.836***	5.25	-0.105**	-2.64
RLGINSTR	3.313*	2.14	0.509*	2.25	-0.000149	-0.04
PCHURCH	0.512	1.63	0.219***	4.80	0.000909	1.19
NOINCOME	172.257***	3.58	6.199	0.88	-0.425***	-3.30
INCOME	9.025***	9.73	0.019	0.14	-0.034***	-4.53
HEDUC	4.124	1.75	0.231	0.67	-0.007	-1.35
AGE	3.702***	7.66	0.316***	4.47	-0.018*	-2.47
SEX	-0.912	-0.07	7.232***	4.13	0.103***	3.70
NONWHITE	-36.259	-1.83	-0.723	-0.25	0.252***	5.34
NKIDS	4.646	1.17	-0.020	-0.03	-0.023*	-2.38
BELIEF	35.256**	2.96	12.591***	7.23	0.039	1.40
NBHD	-25.373	-0.95	-2.725	-0.69	-0.035	-0.56
RAISECA	-5.044	-0.12	-1.068	-0.18	-0.036	-0.37
PCATH	-64.239	-1.62	-11.243**	-1.94	0.025	0.26
(CONSTANT)	-194.256	-4.02	4.122	2.58	1.276	7.64
R-squared	.39		.28		.31	
Cases	555		555		456	

*p ≤ .05 **p ≤ .01 ***p ≤ .001

NOTES:
Coefficients: Unstandardized regression coefficients.
Source: N.O.R.C. American Catholic Survey, 1974.
Sample: All married respondents.
Variable definitions:
AGE = respondent's age.
ATTEND = yearly number of masses attended.
BELIEF = 9-item additive scale of respondent's strength of religious belief.
CONTRIB = yearly contributions to church (excluding Catholic school tuition and contributions).
HEDUC = years of education of family head.
INCOME = yearly income (thousands).
MARSAME = coded 1 if respondent and spouse of same religion.
NBHD = fraction of Catholic neighbors when growing up.
NKIDS = number of preschool or school-age children.
NOINCOME = dummy (1 if income not reported, 0 otherwise).

NONWHITE = dummy (1 if respondent is nonwhite, 0 otherwise).
PCATH = dummy (1 if either parent Catholic, 0 otherwise).
PCHURCH = mean of parents' yearly mass attendance.
RAISECA = dummy (1 if respondent was raised a Catholic, 0 otherwise).
RATIO = time intensity of religious participation - ATTEND/CONTRIB.
RLGINSTR = respondent's religious instruction scale score.
SEX = sex of respondent (1 if female, 0 if male).

II. [NOTE: Even if you think you are a Stata "expert," for this problem you should read the handout on Using Stata for OLS Regression, especially the section on Analyzing Means, Correlations and Standard Deviations.] Download the file *sphrd.dta* from the course web page. As explained in the handout on using Stata for OLS regression, this data set was created using Stata's `corr2data` command based on results published in the 1985 ASR paper, "Ability grouping and contextual determinants of educational expectations in Israel." In that piece, Shavit and Williams examined the effect of ethnicity and other variables on the achievement of Israeli school children. There are two main ethnic groups in Israel: the Ashkenazim - of European birth or extraction - and the Sephardim, most of whose families immigrated to Israel during the early fifties from North Africa, Iraq, and other Mid-eastern countries. Their variables included:

X1 - Ethnicity (`sphrd`) - a dummy variable coded 1 if the respondent or both his parents were born in an Asian or North African country, 0 otherwise

X2 - Parental Education (`pared`) - A scale which ranges from a low of 0 to a high of 1.697

X3 - Scholastic Aptitude (`aptd`) - A composite score based on seven achievement tests.

Y - Grades (`grades`) - Respondent's grade-point average during the first trimester of eighth grade. This scale ranges from a low of 4 to a high of 10.

Analyze these data using Stata and answer the following questions. Begin with the command

```
regress grades sphrd pared aptd
```

and then execute whatever other commands are necessary. Note that NO hand computation is needed. All you have to do is run the analyses in Stata and then interpret the results.

1. What is the metric (unstandardized) coefficient for the effect of `aptd` on `grades`? What is the 99% confidence interval for this effect?

```
. use "https://www3.nd.edu/~rwilliam/statafiles/sphrd.dta"
(Simulation of Shavit-Williams data)

. regress grades sphrd pared aptd, lev(99)

      Source |       SS       df       MS              Number of obs =   10609
-------------+------------------------------           F(  3, 10605) = 3915.88
       Model |  11241.6967     3  3747.23224           Prob > F      =  0.0000
    Residual |  10148.2747 10605  .956933021           R-squared     =  0.5256
-------------+------------------------------           Adj R-squared =  0.5254
       Total |  21389.9714 10608  2.01640002           Root MSE      =  .97823

------------------------------------------------------------------------------
      grades |      Coef.   Std. Err.      t    P>|t|     [99% Conf. Interval]
-------------+----------------------------------------------------------------
       sphrd |   .2117594    .024091     8.79   0.000      .149694    .2738248
       pared |  -.1106953   .0274186    -4.04   0.000    -.1813336    -.040057
        aptd |   .5204228   .0054355    95.75   0.000     .5064195    .5344262
       _cons |   3.755664   .0418421    89.76   0.000     3.647867    3.863462
------------------------------------------------------------------------------
```

The metric coefficient for the effect of aptitude on grades is .5204228. The 99% confidence interval goes from .5064195 to .5344262.

2. What is the standardized coefficient for the effect of `sphrd` on `grades`?

```
. reg, beta

      Source |       SS       df       MS              Number of obs =   10609
-------------+------------------------------           F(  3, 10605) = 3915.88
       Model |  11241.6967     3  3747.23224           Prob > F      =  0.0000
    Residual |  10148.2747 10605  .956933021           R-squared     =  0.5256
-------------+------------------------------           Adj R-squared =  0.5254
       Total |  21389.9714 10608  2.01640002           Root MSE      =  .97823

------------------------------------------------------------------------------
      grades |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       sphrd |   .2117594    .024091     8.79   0.000                 .0745632
       pared |  -.1106953   .0274186    -4.04   0.000                -.035859
        aptd |   .5204228   .0054355    95.75   0.000                 .7733043
       _cons |   3.755664   .0418421    89.76   0.000                       .
------------------------------------------------------------------------------
```

The standardized coefficient for sphrd is .0745632.

3. Test the hypothesis $\beta_{pared} = \beta_{aptd} = 0$. (Remember, it is very easy to do this in Stata.)

```
. test pared aptd

 ( 1)  pared = 0
 ( 2)  aptd = 0

       F(  2, 10605) = 5118.30
            Prob > F =    0.0000
```

The hypothesis that both effects = 0 should clearly be rejected. This is consistent with the t tests, which showed that both effects were significant.

4. What percentage of the respondents are Sephardim?

```
. sum sphrd if !missing(grades, sphrd, pared, aptd)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       sphrd |     10609         .44          .5  -1.339484   2.122911
```

44% are Sephardim. This is one of several ways to get the result; the `if` parameter guarantees that the same listwise deletion that is done on the regression command gets done with the summarize command.

Incidentally, since sphrd in the original data was a dichotomy, it might be tempting just to use a tab command – but it won't work:

```
. tab  sphrd
too many values
r(134);
```

This is because `corr2data` creates a data set that has the same means, variances and covariances as the original data; but it has no way of knowing how the variables were

originally measured, so it just creates continous variables for everything. Be careful how you use data created by `corr2data`; only use it for analyses where the means, correlations and standard deviations are sufficient for the calculations.

5. What are the partial, semipartial, and zero-order (i.e. bivariate) correlations of `pared` with `grades`?

```
. pcorr2  grades sphrd pared aptd

(obs=10609)

Partial and Semipartial correlations of grades with

    Variable |     Partial      SemiP    Partial^2     SemiP^2       Sig.
-------------+------------------------------------------------------------
       sphrd |      0.0850      0.0588      0.0072      0.0035       0.000
       pared |     -0.0392     -0.0270      0.0015      0.0007       0.000
        aptd |      0.6809      0.6404      0.4636      0.4101       0.000
```

The partial and semipartial correlations are -0.0392 and -0.027. Or, if you have Stata 11 or higher, you can just use the built-in `pcorr` command (which StataCorp updated by borrowing code from Rich Williams' `pcorr2` command):

```
. pcorr  grades sphrd pared aptd
(obs=10609)

Partial and semipartial correlations of grades with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2          Value
------------+------------------------------------------------------------------
      sphrd |   0.0850        0.0588       0.0072        0.0035         0.0000
      pared |  -0.0392       -0.0270       0.0015        0.0007         0.0001
       aptd |   0.6809        0.6404       0.4636        0.4101         0.0000

. corr grades sphrd pared aptd
(obs=10609)

             |   grades    sphrd    pared     aptd
-------------+------------------------------------
      grades |   1.0000
       sphrd |  -0.2600   1.0000
       pared |   0.3300  -0.5900   1.0000
        aptd |   0.7200  -0.4600   0.5300   1.0000
```

The bivariate correlation is .33.

6. In their published analyses, Shavit and Williams reported that

$$E(\text{grades}) = .185*\text{sphrd} - .119*\text{pared} + .49*\text{aptd} + 4.057$$

Your results should be close, but not identical to this. Explain what might account for the discrepancy.

The cases used when reporting the correlation matrix, means and standard deviations were probably not the exact same cases used in all parts of the analysis. Differing amounts of missing data in variables can cause the Ns and sample selection to differ.