PLSC 503: Problem Set 4 Solutions

Thad Dunning

Department of Political Science

Yale University

Spring 2010

1 Theoretical/Conceptual Exercises

Question 1: Suppose that you have an $n \times 1$ column vector of observations on a variable Y and an n $\times 1$ column vector of observations on a variable X. You convert each observation for each variable to standard units. Then you run a regression of standardized Y on standardized X. Show that the fitted slope coefficient on standardized X is r, the coefficient of correlation between Y and X.

Solution: Note that if we regress the (unstandardized) Y on an intercept and X, we can express the actual values in terms of the fitted values and the residuals: $Y_i = \hat{a} + \hat{b}X_i + e_i$, where $\hat{a}$ and $\hat{b}$ are the fitted intercept and slope coefficient, respectively. Thus, recalling that $\bar{e} = 0$, we have $\bar{Y} = \hat{a} + \hat{b}\bar{X}$, and so $Y_i - \bar{Y} = \hat{b}(X_i - \bar{X}) + e_i$. (Just subtract the expression for $\bar{Y}$ from the expression for Y_i). Then,

$$\frac{Y_i - \bar{Y}}{S_Y} = \hat{b}\frac{S_X}{S_Y}\frac{X_i - \bar{X}}{S_X} + \frac{e_i}{S_Y}, \tag{1}$$

where we have divided through by the standard deviation S_Y and then multiplied and divided $\hat{b}\frac{X_i - \bar{X}}{S_Y}$ by $\frac{S_X}{S_X}$. Now, we now that the fitted slope coefficient in the unstandardized case is

$$\hat{b} = \frac{\text{cov}(X, Y)}{\text{var}(X)}. \tag{2}$$

So, the fitted coefficient on $\frac{X_i - \bar{X}}{S_X}$ in (1) is

$$\hat{b}\frac{S_X}{S_Y} = \frac{\text{cov}(X, Y)}{\text{var}(X)}\frac{S_X}{S_Y} = \frac{\text{cov}(X, Y)}{(S_X)(S_Y)} = r. \tag{3}$$

The key point: the slope coefficient in a standardized bivariate regression (that is, a regression of standardized Y on standardized X) is just r, the correlation coefficient between X and Y.

Question 2: Exercise 6.C.2 in Freedman (2009), that is, exercise C.2 in Chapter 5.

Solution: Follow the method outlined on pages 83-85 of Freedman (2009), for the Blau and Duncan article. First, the (partitioned) design matrix is here $X = (ME)$, where M is a column vector of (standardized) mass tolerance scores and E is a column vector of (standardized) elite tolerance scores. The column vector Y gives standardized repression scores. We will estimate the model

$$Y = X\beta + \epsilon, \tag{4}$$

where $\beta = (\beta_1 \beta_2)'$ are the path coefficients. So,

$$X'X = n \begin{pmatrix} 1.00 & 0.52 \\ 0.52 & 1.00 \end{pmatrix}, \tag{5}$$

and

$$(X'X)^{-1} = \frac{1}{n(1 - (0.52)^2)} \begin{pmatrix} 1.00 & -0.52 \\ -0.52 & 1.00 \end{pmatrix}. \tag{6}$$

Next,

$$X'Y = n \begin{pmatrix} -0.26 \\ -0.42 \end{pmatrix}. \tag{7}$$

Thus, the path coefficients are

$$
\begin{aligned}
\begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \end{pmatrix} &= (X'X)^{-1}X'Y \\
&= \frac{1}{n(1-(0.52)^2)} \begin{pmatrix} 1.00 & -0.52 \\ -0.52 & 1.00 \end{pmatrix} n \begin{pmatrix} -0.26 \\ -0.42 \end{pmatrix} \\
&= \frac{1}{(1-(0.52)^2)} \begin{pmatrix} -0.26 + (-0.52)(-0.42) \\ (-0.52)(-0.26) + (1.00)(-0.42) \end{pmatrix} \\
&= \begin{pmatrix} -0.057 \\ -0.390 \end{pmatrix}. \qquad (8)
\end{aligned}
$$

The first coefficient matches the −.06 in the path diagram on p. 88 and p. 324 of Freedman (2009), but the second coefficient is not quite the same as −0.35. This could be due to the weighting (see Freedman 2009: 90) or other factors: see http://www.stat.berkeley.edu/users/census/repgibson.pdf.

Question 3: Exercise 6.C.3 in Freedman (2009).

Solution: Let $\hat\delta$ be the vector of residuals after fitting equation (10) on p. 89 of Freedman (2009), and let $\hat\sigma^2$ be the mean square of the residuals. Let r_{ME} be the correlation between the columns of X. Rearranging the analogue of equation (8) on Freedman (2009: 85) gives

$$
\begin{aligned}
\hat\sigma^2 &= 1 - \hat\beta_1^{\,2} - \hat\beta_2^{\,2} - 2\hat\beta_1\hat\beta_2 r_{ME} \\
\hat\sigma^2 &= 1 - (-0.057)^2 - (-0.390)^2 - 2(-0.057)(-0.390)(0.52) \\
&= 0.822. \qquad (9)
\end{aligned}
$$

However, to estimate the variance of the error term δ, we should have divided by the degrees of freedom $n-p$ instead of n (see e.g. the last line of equation 7 on p. 85 of Freedman). We can adjust by multiplying $\hat{\sigma}^2$ by $\frac{n}{n-p}$; here, n is 36 and $p = 3$, so $\frac{n}{n-p} = 1.091$. The reason that $p = 3$ instead of 2 is that there is implicitly an intercept: we standardized the variables by subtracting the mean from each (unstandardized) observation.

Thus,

$$1.091 * 0.822 = 0.897 \tag{10}$$

estimates the variance of δ, and the estimated SD of δ is

$$\sqrt{0.897} = 0.947. \tag{11}$$

Question 4: Exercise 6.C.4 in Freedman (2009). Conduct a test of the hypothesis that $\beta_2 = \beta_1$.

Solution: The estimated variance-covariance matrix of $\hat{\beta}$ given X is, as usual,

$$\widehat{\text{cov}}(\hat{\beta}|X) = \hat{\sigma}^2(X'X)^{-1}. \tag{12}$$

We just calculated $\hat{\sigma}^2 = 0.897$, and $(X'X)^{-1}$ in equation (6) reduces to

$$\begin{pmatrix} 0.0381 & -0.0198 \\ -0.0198 & 0.038 \end{pmatrix}. \tag{13}$$

Thus,

$$\begin{aligned} \widehat{\text{cov}}(\hat{\beta}|X) &= 0.897\begin{pmatrix} 0.0381 & -0.0198 \\ -0.0198 & 0.0381 \end{pmatrix} \\ &= \begin{pmatrix} 0.0342 & -0.0178 \\ -0.0178 & 0.0342 \end{pmatrix} \end{aligned} \tag{14}$$

The standard errors for the path coefficients are the square roots of the diagonal elements of (14), that is, $\sqrt{0.0342} = 0.185$. Notice therefore that $\hat{\beta}_1 = -0.057$ is not significant, while $\hat{\beta}_2 = -0.390$ is significant, since $|\frac{-0.390}{0.185}| \doteq 2.11$, which exceeds the critical threshold of about 2.

Now, the difference $\hat{\beta}_2 - \hat{\beta}_1 = -0.390 - (-0.057) = -0.333$, and the SE for the difference is

$$\sqrt{\text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_1) - 2\text{cov}(\hat{\beta}_2, \hat{\beta}_1)} = \sqrt{0.0342 + 0.0342 - 2(-0.0178)} = 0.323. \qquad (15)$$

To conduct a test of the hypothesis that $\beta_2 = \beta_1$, we take the absolute value of the difference, divided by the estimated standard error of the difference:

$$t = |\frac{-0.333}{0.323}| = 1.03. \qquad (16)$$

Since the t-statistic fails to exceed 2 (or even 1.96), we cannot reject the null hypothesis that $\beta_2 = \beta_1$. That is, the difference in the coefficients is not significantly different from zero.

Question 5: Solution/commentary omitted.

2 Computer exercises

1. Solution omitted – see the hint in the problem set, or ask Mario or me if you have questions.