Public Policy 529
Fall 2023 Problem Set #9

Francisco Brady

2023-11-29

Due on Wednesday, November 29th

1. Use the `world` dataset for the following:

(a) Using your software, find the Pearson's r correlation coefficients between democracy (`dem_score14`), life expectancy (`lifeex_total`), and the level of poverty (`unpovnpl`). Make sure you have your software report statistical significance (see lecture slide 24 for commands).
Report the resulting correlation matrix in your answers.

The variable `dem_score` is a rating of a country's degree of democracy on a scale that goes from 1 to 10; `lifeex_total` is life expectancy at birth in a country; `unpovnpl` is the percentage of a country's population that is below the national poverty line.

```r
# note: this is the code for question 1a:
# using the code from class does not supply p-values
cor(world[c("dem_score14", "lifeex_total", "unpovnpl")],
    use = "pairwise.complete.obs")
```

```
##              dem_score14 lifeex_total   unpovnpl
## dem_score14    1.0000000    0.5485013 -0.2880678
## lifeex_total   0.5485013    1.0000000 -0.5264596
## unpovnpl      -0.2880678   -0.5264596  1.0000000
```

```r
# using the Hmisc package for the rcorr function and the broom package
# we can output the p-values
# library(broom)
world_mat <- as.matrix(world[c("dem_score14", "lifeex_total", "unpovnpl")])
world_cor <- Hmisc::rcorr(world_mat, type = 'pearson')
kable(as.data.frame(tidy(world_cor)),
      col.names = c('x', 'y', 'Cor.', 'n', 'p-value'),
  caption = 'With p-values')
```

With p-values

x	y	Cor.	n	p-value
lifeex_total	dem_score14	0.5485013	164	0.000000
unpovnpl	dem_score14	-0.2880678	72	0.014135

x	y	Cor.	n	p-value
unpovnpl	lifeex_total	-0.5264596	72	0.000002

(b) Interpret the correlation coefficients for each pairing of the three variables (not counting each variable with itself).

- `dem_score14` has a moderately-positive correlation with `lifeex_total`. When democracy scores are higher in a country, total life expectancy tends to be higher.

- `dem_score14` has a weakly-negative correlation with `unpovnpl`. When democracy scores are higher in a country, the percentage of the population below the poverty line tends to be lower.

- `lifeex_total` has a moderately-negative correlation with `unpovnpl`. When life expectancy in a country is lower, the percentage of the population below the poverty line tends to be higher.

(c) If the sample size for the correlation between `dem_score14` and `unpovnpl` were 82, what would be the t-statistic and p-value for the correlation coefficient?
Since r and n are given, we can calculate the standard error and the test statistic (against the t-distribution):

$$\begin{aligned} se_r &= \sqrt{\frac{1-r^2}{n-2}} \\ se_r &= \sqrt{\frac{1-(-0.2880678)^2}{82-2}} \\ se_r &= \sqrt{\frac{0.9170169}{80}} \\ se_r &= 0.1070641 \end{aligned}$$

Plugging this in to calculate the t-statistic:

$$\begin{aligned} t &= \frac{r}{se_r} \\ t &= \frac{-0.2880678}{0.1070641} \\ t &= -2.690611 \end{aligned}$$

To find our critical value, we use the t-distribution table and find the desired confidence interval. In this case we are not given an alpha so we can choose $\alpha = 0.05$. Along the 80 degrees of freedom row, our critical value is: 1.990. The test statistic is between 2.639 and 3.195 in the 80 *df* row, so we can say our p-value is between $t_{.002}$ and $t_{.01}$.

2. This question uses the `world` dataset. Let's use bivariate linear regression to estimate the linear relationship between `educ_f_none` (dependent variable) and `democ11` (independent variable). The variable `educ_f_none` is the percentage of females in a country with no schooling. `democ11` is an 11-point scale of democracy, which runs from 0 ("not at all democratic") to 10 ("highly democratic"). Note: we are treating `democ11` as an interval-level variable for this analysis.

```r
# note: code for question 2a:
model <- lm(educ_f_none ~ democ11, data = world)
summary(model)
```

(a) Estimate the regression and report the output. See lecture slides or help documents for appropriate commands.

```
##
## Call:
## lm(formula = educ_f_none ~ democ11, data = world)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.860 -10.737  -7.958   5.850  68.870
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept)  35.8942     4.2065   8.533 0.0000000000000598 ***
## democ11      -2.5209     0.5721  -4.407 0.0000234163957921 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.37 on 117 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.1423, Adjusted R-squared:  0.135
## F-statistic: 19.42 on 1 and 117 DF,  p-value: 0.00002342
```

(b) What is the substantive meaning of the estimated intercept? In other words, what does it tell us about the predicted percentage of females with no schooling?

The intercept is 35.89, which implies that in a country where the democracy score is 0 ("not at all democratic"), the predicted percent of women with no schooling is 35.89%.

(c) What is the coefficient on `democ11`? Is this coefficient statistically significant at the .05 level of significance? How do you know?

The coefficient on democracy score is -2.52. The coefficient is statistically significant at the .05 level. We know this because the model outputs a p-value of 0.0000234163957921, which is the probability of this result if the null hypothesis was true ($H_0 : \beta_1 = 0$). This p-value is lower than the .05 level of significance.

(d) Interpret the substantive meaning of the coefficient on `democ11`. In other words, what does it tell us about the relationship between the democracy scale and female schooling? Remember that regression coefficients give us information about both the direction and magnitude of this relationship.

For each additional increase of one unit along the democracy scale, is associated with a 2.52 percentage point decrease in the percentage of women that have no education. There is a negative association between the two variables.

(e) If a country were a 4 on the democracy scale, what would we predict to be the percentage of females with no schooling?

We can calculate this using the formula given by the output:

$$\begin{aligned} educ_{f_{none}} &= 35.8942 + -2.52 \cdot \text{democ11} \\ educ_{f_{none}} &= 35.8942 + -2.52(4) \\ educ_{f_{none}} &= 35.8942 + (-10.08) \\ educ_{f_{none}} &= 25.8142 \end{aligned}$$

(f) Country A is a 7 on the democracy scale; Country B is a 2. What is the predicted difference in female schooling?
Using the function, we can use the score for the two countries and subtract the calculated values to find the predicted difference in women with no schooling:

$$\begin{aligned} educ_{f_{none_A}} &= 35.8942 + -2.52(7) \\ educ_{f_{none_A}} &= 18.2542 \\ educ_{f_{none_B}} &= 35.8942 + -2.52(2) \\ educ_{f_{none_B}} &= 30.8542 \\ diff = educ_{f_{none_A}} &- educ_{f_{none_B}} \\ diff &= 18.2542 - 30.8542 \\ diff &= -12.6 \end{aligned}$$

(g) Interpret the R^2 statistic for this regression.
The R^2 in the output is 0.1423, which implies that the democracy score "explains" or accounts for only around 14% of the variation in the percentage of women with no education in each country.

(h) What is the size of the typical difference between the predicted percentage of females with no schooling and the actual percentage of females with no schooling (i.e. the standard error of the estimate)? Stata calls this the Root MSE; R calls it the residual standard error.
The reported residual standard error is 22.37. This is the standard deviation of the entire model, which in this case is one variable, so the interpretation is that on average the distance between the value of `educ_f_none` predicted by `democ11` and the actual value is 22.37 percentage points.