PS0700
Basic Statistical Methods: Multivariate Analysis

Political Science Research Methods

Professor Steven Finkel

Fall Semester 2022

Week 14

Why Multivariate Analysis?

- We need to control for third, fourth, etc. "Z" variables so that we get the "true" (unbiased) effect of the primary independent variable of interest on the dependent variable
  - Is X truly related to Y or is the relationship "spurious"?
  - Is a policy intervention truly responsible for some outcome, or is it because the people or places exposed to the intervention already differed on some important variable that produced the outcome (i.e., the selection problem in quasi-experimental research)
  - In non-experimental research, we cannot be sure without controlling for as many Z variables as we plausibly can (and even then, we cannot be 100% "sure" because of unmeasured variables that may be relevant!)

- With multivariate analysis, we obtain a better understanding of *all* (or at least more of) the factors that explain the dependent variable
  - No relationship in social or policy sciences is mono-causal, so multivariate explanations are more likely to be correct, i.e., predict the DV better (increase "R-squared")
  - Introducing additional variables may help clarify which ones are the most important predictors of Y
  - Introducing additional variables may help clarify the conditions under which each one has strongest effects on Y

Hypothetical Example with Cross-Tabs

- Did participation in the job training program lead to increased wages, compared to control group that did not participate in the program?
- Bivariate Relationship
  - Percentage Difference=22%
  - Chi-Square=45.7 with 1 df, p<.001

	Control	Treatment	Total
Low Wage Increase	200 50%	140 28%	340 38%
High Wage Increase	200 50%	360 72%	560 62%
Total	400 44%	500 56%	

Alternative Causal Specification:

- People who participated in the job training program were more highly educated, and more educated people generally increased their earnings (wages) more so than less educated people. So it is not the program that caused the increase in wages, rather the pre-existing level of education among people who were exposed to the training
- If true, the bivariate relationship would be "spurious," due to their joint relationship with Z (education)

$$\begin{tikzcd}
\text{Education} \arrow[r] \arrow[dr] & \text{Job Training} \\
& \text{Wage Increases}
\end{tikzcd}$$

How to Test this Alternative Hypothesis?

- We “control” for education by examining the bivariate relationship between Job Training and Wages among people who are poorly educated and among people who are highly educated
- That is, we “hold education constant” and see whether the original relationship is maintained once the controls are introduced
- In cross-tabs, these controls are done “manually” by separating the sample into “low” and “high” education groups (or low, medium, high) and conducting the cross-tab analysis (percent differences, chi-square) for each of the sub-samples
- Same with t-tests: we separate the sample into different educational groups and conduct the t-test analysis for each of those sub-samples
- In regression, the controls are done “statistically” by adjusting the calculation of the X-Y slope to take into account their mutual correlation with Z.
- The logic of multivariate analysis – regardless of the kinds of variables you have -- is the same!!

LOW EDUCATION	Control	Treatment	Total
Low Wage Increase	180 60%	60 60%	240 60%
High Wage Increase	120 40%	40 40%	160 40%
Total	300 75%	100 25%	400

HIGH EDUCATION	Control	Treatment	Total
Low Wage Increase	20 20%	80 20%	100 20%
High Wage Increase	80 80%	320 80%	400 80%
Total	100 20%	400 80%	500

This is "Perfect" Spuriousness!!

- Among people with low education, 40% of the control group had high wage increases, and 40% of the treatment group also had high wage increases. So no effect of job training among people with low education (i.e. percentage difference) = 0
- Among people with high education, 80% of the control group had high wage increases, and 80% of the treatment group also had high wage increases. So no effect of job training among people with high education (i.e. percentage difference) = 0)
- Since there is no effect of job training on wages among people with either low or high education, we conclude there is "no effect of job training on wages, controlling for education"

How did this happen?

1. ***Z was related to Y***: People who had high levels of education were more likely to have high wage increases, generally speaking. We know this because 400 of the 500 highly educated people (80%) were in the high wage increase group, and only 160 of the 400 low educated people (40%) were in the high wage increase group.
2. ***Z was related to X***: People who had high level of education were more likely to seek out job training, generally speaking. We know this because 400 of the 500 highly educated people (80%) were in the treatment group, while only 100 of the 400 (25%) of the low educated people were in the treatment group.
3. ***The effects of #1 and #2 above were enough to wipe out the observed bivariate relationship between X and Y***

Other Possible Outcomes of Multivariate Analysis

2. Controlling for Z, you find that the original relationship between X and Y is **weaker but still exists**

Conclusion: X and Z are both important in explaining Y

Hypothetical Example: People with more prior work experience tend to participate more in the job training program. Prior work experience helps people increase their wages over time, and so does participation in the training program. But controlling for prior work experience, training matters less than it originally appeared.

Empirical Pattern: Bivariate Effect of X on Y is weaker but still significant at all levels of Z; the combination of X and Z gives greater predictive accuracy than either variable by itself

Other Possible Outcomes of Multivariate Analysis

3. Controlling for Z, the original relationship between X and Y is **unchanged**.

Example: The effect of participation in job training program on wage increases is the same for married and unmarried persons

Empirical Pattern: X affects Y, regardless of the level of Z, and the effect is the same as it was in the bivariate analysis

Conclusion: X matters, Z ***may or may not*** matter in explaining Y (you have to check this out from the pattern of results, it is not inherently one way or the other)

Other Possible Outcomes of Multivariate Analysis

4. Controlling for Z, you find that X affects Y at some levels of Z but not others. So X and Y are said to have a ***"conditional relationship"***, depending on the level of Z

Empirical Pattern: Bivariate effect of X on Y is diminished or wiped out at one level of Z but is stronger at another level of Z

This is a VERY common pattern in empirical policy research, and it is also of much practical use to know whether and where policy interventions have greater or weaker impacts

Hypothetical Example: Do the effects of a job training program on wages depend on the age of the employee? Do younger or older employees benefit more from the program, or are the effects the same, regardless of age?

```
        Welch Two Sample t-test

data:  salchange by treatment
t = -4.4515, df = 432.69, p-value = 1.086e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -616.4933 -238.8391
sample estimates:
mean in group 0 mean in group 1
       1442.631        1870.298
```

Overall relationship among all employees: Participants in the job training program show a $428 greater increase than non-participants; t-test is statistically significant

```
        Welch Two Sample t-test

data:  salchange by treatment
t = -1.3983, df = 260.12, p-value = 0.1632
alternative hypothesis: true difference in me
95 percent confidence interval:
 -423.21151   71.73862
sample estimates:
mean in group 0 mean in group 1
       1862.320        2038.056
```

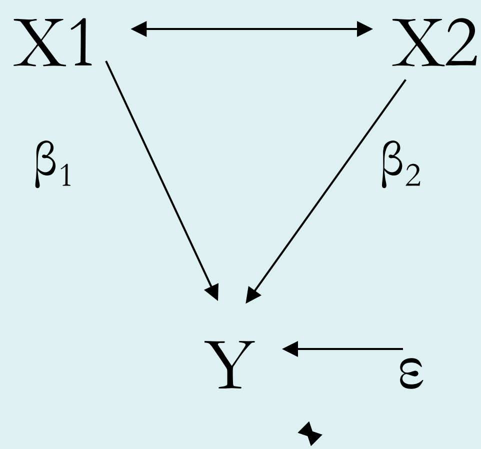
```
        Welch Two Sample t-test

data:  salchange by treatment
t = -6.1051, df = 113.57, p-value = 1.468e-08
alternative hypothesis: true difference in mea
95 percent confidence interval:
 -948.7657 -483.8797
sample estimates:
mean in group 0 mean in group 1
       866.1363        1582.4590
```

Among younger employees (less than 35), participants in the job training program show a $174 greater increase; t-test is *not* statistically significant

Among older employees (greater than 35), participants in the job training program show a $716 greater increase; t-test is statistically significant. **Overall, there is a conditional relationship between job training and wages, depending on age!!**

Multiple Regression Analysis

- Same logic of multivariate analysis in general: We introduce Z (or what could be called "X2") into the process to see whether X1 is truly related to Y, once Z (X2) is controlled, and to see whether X1 and X2, taken together, provide a better explanation of Y than either by itself



Estimation of Multiple Regression Coefficients

- Logic: Take out the part of X1 that is related to X2, and take out the part of Y that is related to X2, and then regress what is left from X1 on what is left from Y!
- This is then the effect of X1 on Y with no influence of X2 on the process at all, or, "controlling for X2," or, "holding X2 constant"
- These effects are called "partial slopes"

Formula

- Bivariate Slope: $\beta = r_{yx1}(\frac{SD_y}{SD_{x1}})$
- Multivariate Slope:

$$\beta_1(x_1) = (\frac{SD_y}{SD_{x1}})(\frac{r_{yx1} - r_{yx2} r_{x1x2}}{1 - r_{x1x2}^2})$$

$$\beta_1(x_2) = (\frac{SD_y}{SD_{x2}})(\frac{r_{yx2} - r_{yx1} r_{x1x2}}{1 - r_{x1x2}^2})$$

- What is the difference? Multivariate slope subtracts out the joint correlation of X1 and Y with X2!! That is what it means to "control" for X2 (or to control for X1 in the equation for β_2)!
- If all variables are positively related with each other, the multivariate slope will be *smaller* than the bivariate slope

Example: Job Training and Wage Increases

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1442.63      73.85  19.535  < 2e-16 ***
treatment     427.67     100.10   4.273 2.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1085 on 472 degrees of freedom
Multiple R-squared:  0.03723,	Adjusted R-squared:  0.0352 
F-statistic: 18.25 on 1 and 472 DF,  p-value: 2.339e-05
```

Bivariate “Slope” of Participation in Job Training is 427.67. Since treatment has two values – 0 for non-participants, and 1 for participants – this means that the average salary change for non-participants is predicted to be $1442.63, and the average salary change for participants is predicted to be

(1442.63+427.67=$1870.3) per month. This effect is statistically significant.

R-squared very weak, though, at only .04, so training is not *strongly* associated even at the bivariate level, though it is statistically significant

Example: Job Training and Wage Increases

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1442.63      73.85  19.535  < 2e-16 ***
treatment     427.67     100.10   4.273 2.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1085 on 472 degrees of freedom
Multiple R-squared:  0.03723,   Adjusted R-squared:  0.0352
F-statistic: 18.25 on 1 and 472 DF,  p-value: 2.339e-05
```

Bivariate "Slope" of Participation in Job Training is 427.67. Since treatment has two values – 0 for non-participants, and 1 for participants – this means that the average salary change for non-participants is predicted to be \$1442.63, and the average salary change for participants is predicted to be

(1442.63+427.67=\$1870.3) per month. This effect is statistically significant.

R-squared very weak, though, at only .04, so training is not *strongly* associated even at the bivariate level, though it is statistically significant

Example: Job Training and Wage Increases

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1442.63      73.85  19.535  < 2e-16 ***
treatment     427.67     100.10   4.273 2.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1085 on 472 degrees of freedom
Multiple R-squared:  0.03723,	Adjusted R-squared:  0.0352
F-statistic: 18.25 on 1 and 472 DF,  p-value: 2.339e-05
```

Bivariate "Slope" of Participation in Job Training is 427.67. Since treatment has two values – 0 for non-participants, and 1 for participants – this means that the average salary change for non-participants is predicted to be $1442.63, and the average salary change for participants is predicted to be

(1442.63+427.67=$1870.3) per month. This effect is statistically significant.

R-squared very weak, though, at only .04, so training is not *strongly* associated even at the bivariate level, though it is statistically significant

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -955.42     211.03  -4.527 7.57e-06 ***
treatment       28.35      93.95   0.302    0.763    
edlevel        193.85      16.24  11.940  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 951.9 on 471 degrees of freedom
Multiple R-squared:  0.2609,	Adjusted R-squared:  0.2578 
F-statistic: 83.14 on 2 and 471 DF,  p-value: < 2.2e-16
```

- Controlling for education, the effect of job training on wage increases is only $28.40, with the effect no longer being statistically significant at the .05 level.
- Education has a strong and significant effect on wage increases, controlling for training: for every additional year of education, the individual is predicted to increase wages by nearly $200 per month.
- What happened? Same as in the cross-tab example: Educated people tended to get trained (r=.36), and educated people increased on wages anyway, regardless of being trained or not. The original training→wages relationship was s***purious*** due to the omission of Education in the bivariate model
- R-squared now equal .26 compared to .04 in bivariate model, so model as a whole is better with education as a predictor

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -955.42     211.03  -4.527 7.57e-06 ***
treatment      28.35      93.95   0.302    0.763    
edlevel       193.85      16.24  11.940  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 951.9 on 471 degrees of freedom
Multiple R-squared:  0.2609,	Adjusted R-squared:  0.2578 
F-statistic: 83.14 on 2 and 471 DF,  p-value: < 2.2e-16
```

- Controlling for education, the effect of job training on wage increases is only \$28.4, with the effect no longer being statistically significant at the .05 level.
- Education has a strong and significant effect on wage increases, controlling for training: for every additional year of education, the individual is predicted to increase wages by nearly \$200 per month.
- What happened? Same as in the cross-tab example: Educated people tended to get trained (r=.36), and educated people increased on wages anyway, regardless of being trained or not. The original training→wages relationship was s***purious*** due to the omission of Education in the bivariate model
- R-squared now equal .26 compared to .04 in bivariate model, so the model as a whole is better with education as a predictor

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -955.42     211.03  -4.527 7.57e-06 ***
treatment      28.35      93.95   0.302    0.763
edlevel       193.85      16.24  11.940  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 951.9 on 471 degrees of freedom
Multiple R-squared:  0.2609,	Adjusted R-squared:  0.2578
F-statistic: 83.14 on 2 and 471 DF,  p-value: < 2.2e-16
```

- Controlling for education, the effect of job training on wage increases is only \$28.4, with the effect no longer being statistically significant at the .05 level.
- Education has a strong and significant effect on wage increases, controlling for training: for every additional year of education, the individual is predicted to increase wages by nearly \$200 per month.
- What happened? Same as in the cross-tab example: Educated people tended to get trained (r=.36), and educated people increased on wages anyway, regardless of being trained or not. The original training→wages relationship was s***purious*** due to the omission of Education in the bivariate model
- R-squared now equal .26 compared to .04 in bivariate model, so model as a whole is better with education as a predictor