Path Models

Math 463, Spring 2017, University of Oregon

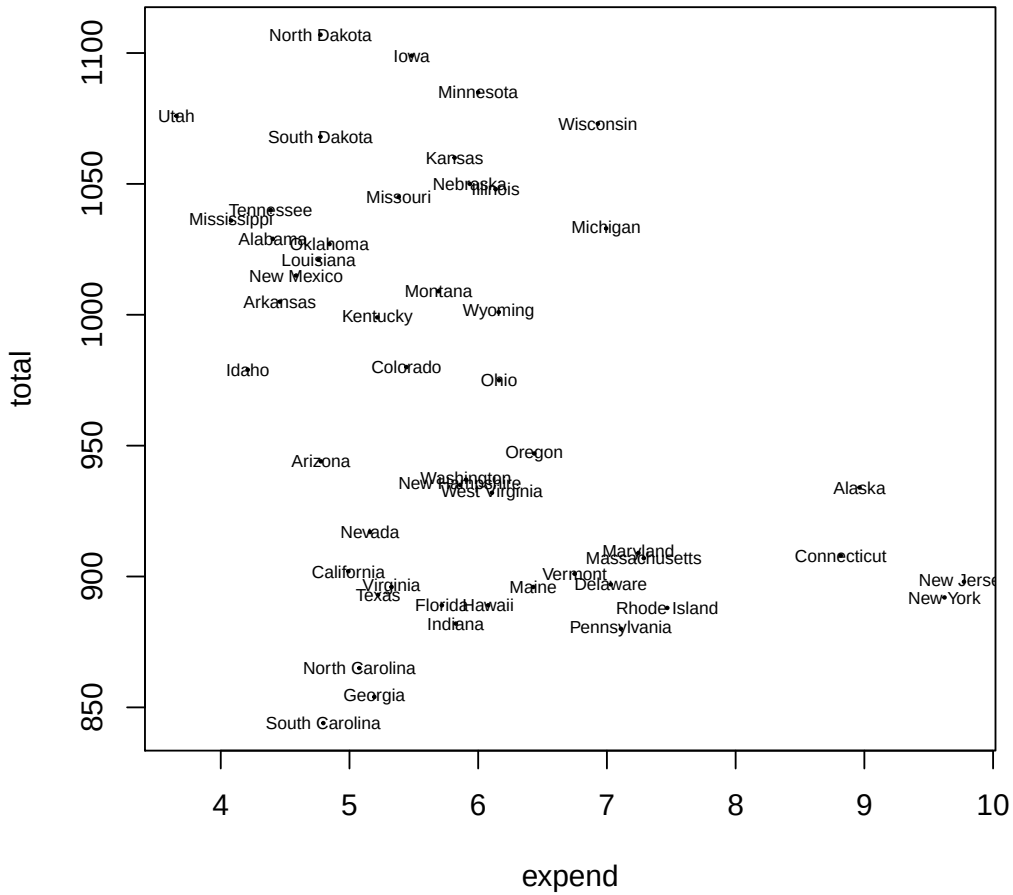
David A. Levin

University of Oregon

May 15, 2017

SAT scores vs. expenditure

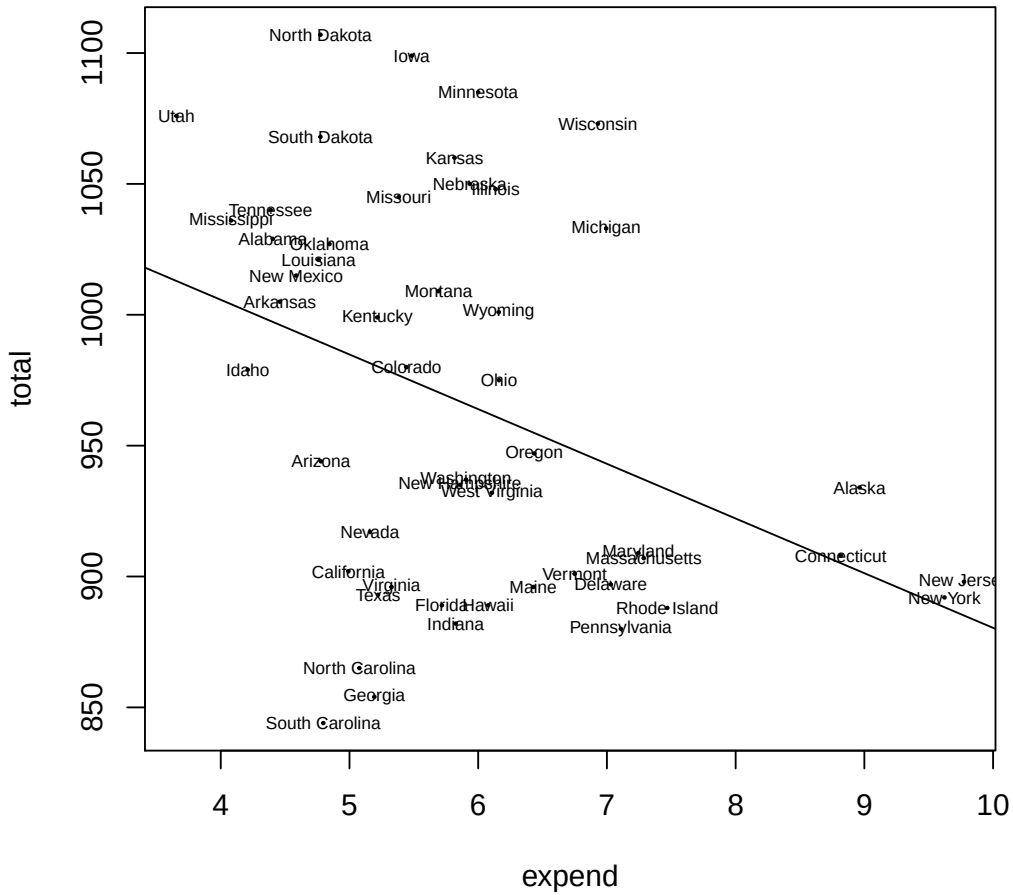
What is the **effect** of spending on performance?



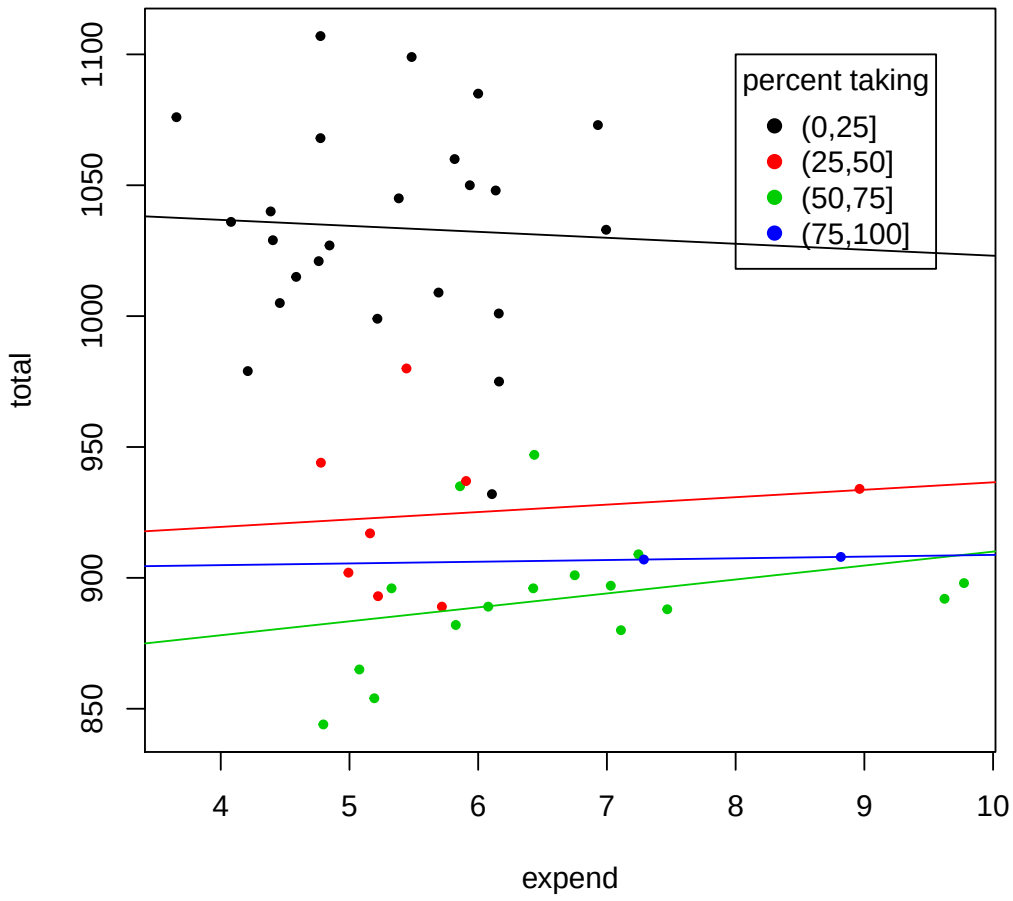
Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars).
Source: "Getting What You Pay For: The Debate Over Equity in Public School Expenditures", D. Guber, Journal of Statistics Education, 1999

SAT scores vs. expenditure

What is the **effect** of spending on performance?



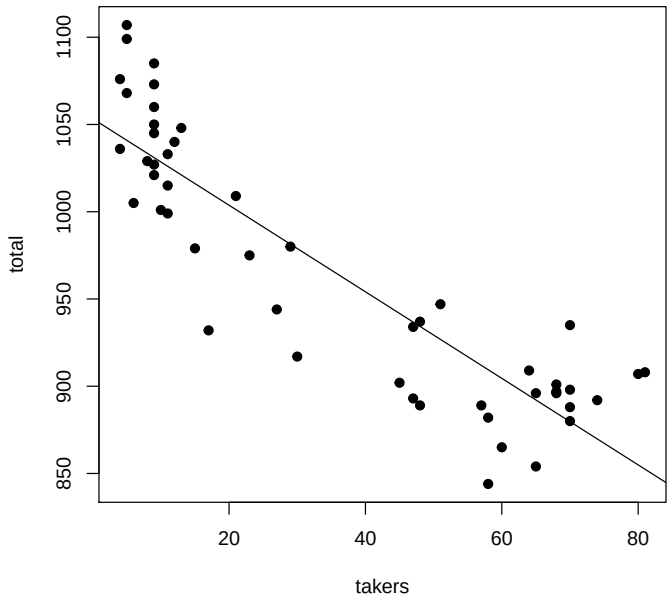
Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars).
Source: "Getting What You Pay For: The Debate Over Equity in Public School Expenditures", D. Guber, Journal of Statistics Education, 1999

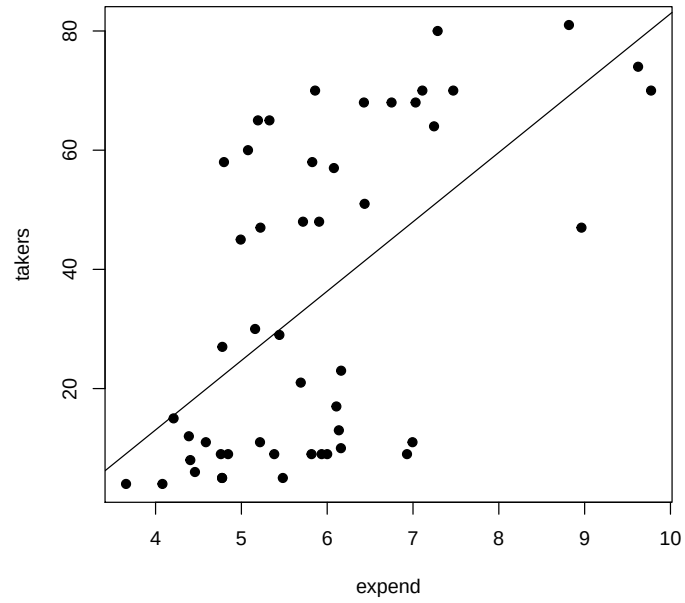

percent taking
(0,25]
(25,50]
(50,75]
(75,100]
total
850
900
950
1000
1050
1100
4
5
6
7
8
9
10
expend

Data

State	expend	ratio	salary	takers	verbal	math	total
Alabama	4.405	17.2	31.144	8	491	538	1029
⋮							
Oregon	6.436	19.9	38.555	51	448	499	947
⋮							
Wyoming	6.16	14.9	31.285	10	476	525	1001

variable name	description
expend	Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)
ratio	Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
salary	Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)
takers	Percentage of all eligible students taking the SAT, 1994-95
verbal	Average verbal SAT score, 1994-95
math	Average math SAT score, 1994-95
total	Average total score on the SAT, 1994-95


total
1100
1050
1000
950
900
850
20
40
60
80
takers


takers
80
60
40
20
4
5
6
7
8
9
10
expend

Regression after standardization

Common in social sciences to center and scale all variables so that they have (sample) mean and (sample) sd 0 and 1, respectively

```
> library(xtable)
> sata = data.frame(scale(
+   read.csv("~/Dropbox/COURSES/MATH463_S17/DATA/sat.csv",
+                                    row.names=1)))
> xtable(cor(sata))
```

	expend	ratio	salary	takers	verbal	math	total
expend	1.00	-0.37	0.87	0.59	-0.41	-0.35	-0.38
ratio	-0.37	1.00	-0.00	-0.21	0.06	0.10	0.08
salary	0.87	-0.00	1.00	0.62	-0.48	-0.40	-0.44
takers	0.59	-0.21	0.62	1.00	-0.89	-0.87	-0.89
verbal	-0.41	0.06	-0.48	-0.89	1.00	0.97	0.99
math	-0.35	0.10	-0.40	-0.87	0.97	1.00	0.99
total	-0.38	0.08	-0.44	-0.89	0.99	0.99	1.00

We imagine two regression equations:

$$\text{takers}_i = a \cdot \text{expend}_i + \varepsilon_i$$

$$\text{total}_i = b \cdot \text{expend}_i + c \cdot \text{takers}_i + \delta_i$$

- Note that if we leave out takers then we have

$$\text{total}_i = b \cdot \text{expend}_i + \gamma_i \,,$$

  where $\gamma_i = c \cdot \text{takers}_i + \delta_i$, and it is not the case that γ_i and expend_i are independent!
- Our estimate of b is then biased. This is *omitted variable* bias.
- In the equations above, we have omitted intercepts. Why?
-

```
> f1 = lm(total~expend+takers-1, data=sata)
> xtable(summary(f1)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
expend	0.22	0.08	2.94	0.01
takers	-1.02	0.08	-13.39	0.00

$\hat{\sigma} = 0.429$

```
> f2 = lm(takers~expend-1, data=sata)
> xtable(summary(f2)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
expend	0.59	0.12	5.15	0.00

$\hat{\sigma} = 0.805$

Dr. A So you see, Dr. Braithwaite, if expenditure goes up by one unit, then takers goes up by 0.59 units.

Dr. B Quite.

Dr. A Furthermore, if expenditure goes up by one unit with takers held fixed, then sat goes up by 0.22 units. This is the direct effect of expenditure on sat. ["Held fixed" means, kept the same; the "indirect effect" is through takers.]

Dr. B But Dr. Arbuthnot, you just told me that if expenditure goes up by one unit, then takers will go up by 0.59 units.

Dr. A Moreover, if takers goes up by one unit with expenditure held fixed, the change in takers makes sat go down by −1.03 units. The effect of takers on sat is −1.03.

Dr. B Dr. Arbuthnot, hello, why would takers go up unless expenditure goes up? "Effects"? "Makes"? How did you get into causation?? And what about my first point?!?

Important points

- Need covariates to be independent of "errors" to obtain unbiased estimates via OLS.
- What about question in HW about the errors in smsa modelling?
- Without experimentation, we can only *assume* that regression equations are *structural*, i.e. are obtained via *response schedules*.
- Randomization imposes independence of randomized variables and errors. We *assume* that nature randomizes the values of the covariates independently of the error.

Figure 1. Path model. Stratification, US, 1962.

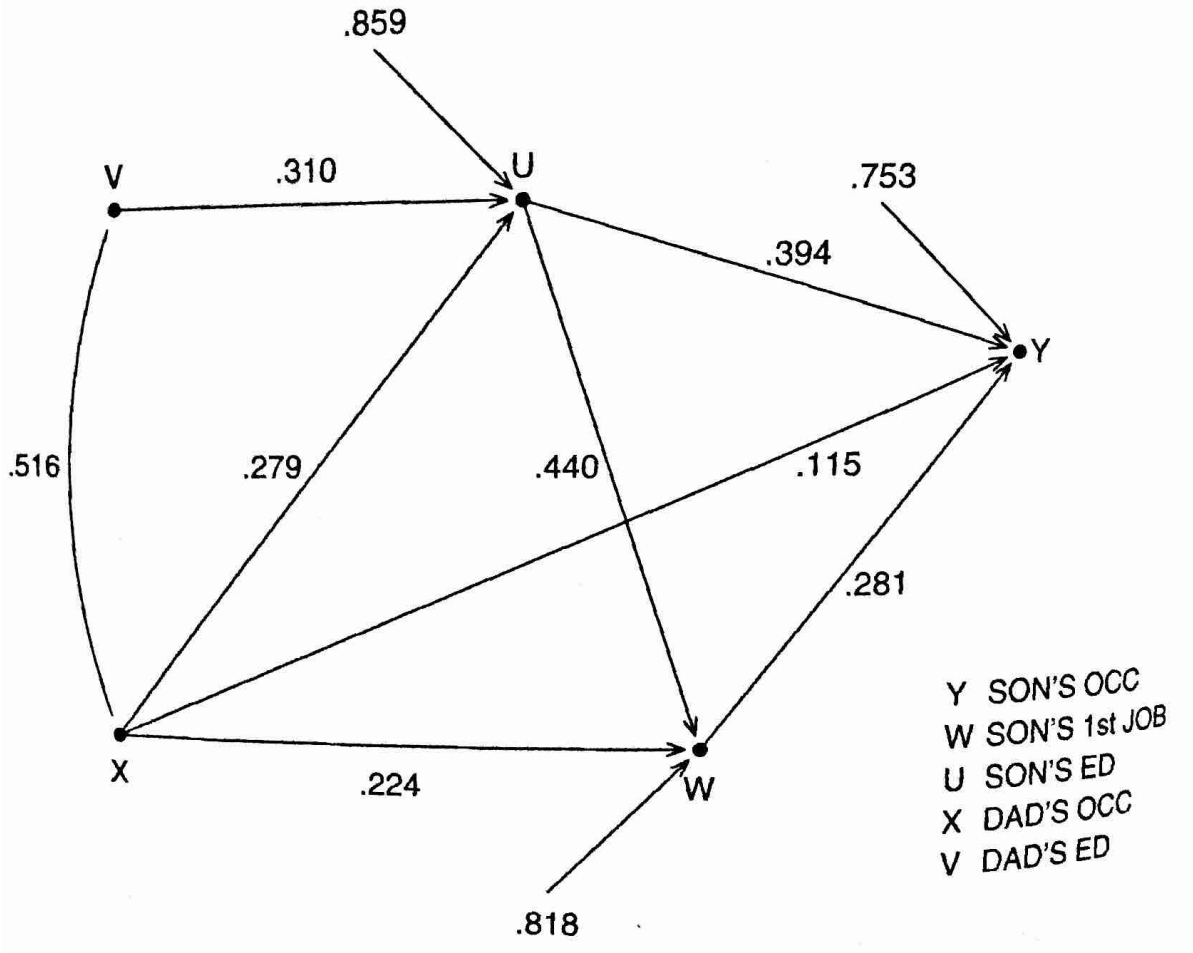


Table 1. Correlation matrix for variables in Blau and Duncan's path model.

		Y Son's occ	W Son's 1st job	U Son's ed	X Dad's occ	V Dad's ed
Y	Son's occ	1.000	.541	.596	.405	.322
W	Son's 1st job	.541	1.000	.538	.417	.332
U	Son's ed	.596	.538	1.000	.438	.453
X	Dad's occ	.405	.417	.438	1.000	.516
V	Dad's ed	.322	.332	.453	.516	1.000

$$W = cU + dX + \epsilon, \tag{2}$$

$$Y = eU + fX + gW + \eta. \tag{3}$$

Another equation

$$U_i = aX_i + bV_i + \delta_i$$

- These path models are often used to tease apart causal relationships.
- Let us consider the simplest model:

$$Y_i = \rho X_i + \varepsilon_i \tag{1}$$

  where X_i, Y_i are both mean zero with standard deviation 1, and ε_i is independent of X_i.
- We know that OLS gives conditionally unbiased estimates of ρ in this case.
- We are sometimes warned that ρ does not measure the "effect of X on Y", but only correlation.
- But doesn't (1) itself imply that Y_i is determined jointly by X_i and ε_i? How is it possible that (1) hold, but Y_i not be "caused" by X_i?

- Suppose that $\boldsymbol{Z} = (Z_1, Z_2)$ is $N(0, I_2)$. If $\boldsymbol{A}$ is a 2×2 matrix, then $\boldsymbol{AZ}$ is $N(0, \boldsymbol{A}'\boldsymbol{A})$.
- Given a symmetric matrix Σ, can we find $\boldsymbol{A}$ with $\boldsymbol{A}'\boldsymbol{A} = \Sigma$?
- Spectral Theorem: $\Sigma = \boldsymbol{U}'\Lambda\boldsymbol{U}$, where $\boldsymbol{U}$ is orthonormal and Λ is diagonal. The entries of Λ are the eigenvalues corresponding to the eigenvectors $\boldsymbol{u}_i$. Then $\boldsymbol{A} = \boldsymbol{U}'\sqrt{\Lambda}\boldsymbol{U}$ is the "square-root" of Σ.
- Let $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The eigenvalues are solutions to

$$0 = \det(\Sigma - \lambda I) = (1 - \lambda)^2 - \rho^2 ,$$

so $\lambda = 1 \pm \rho$. Solving the eigenvector equation shows that

$$\boldsymbol{U} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

- This shows that if $a = \sqrt{1-\rho}$ and $b = \sqrt{1+\rho}$, then

$$\Sigma^{1/2} = \begin{bmatrix} (a+b)/2 & (a-b)/2 \\ (a-b)/2 & (a+b)/2 \end{bmatrix}$$

- Thus, letting

$$X = [(a+b)/2]Z_1 + [(a-b)/2]Z_2, \quad Y = [(a-b)/2]Z_1 + [(a+b)/2]Z_2$$

  we have

$$Y = \rho X + \varepsilon\,,$$

  where $\varepsilon = (Y - \rho X)$.
- Note

$$\mathrm{Cov}(X, \varepsilon) = \rho - \rho = 0\,,$$

  and since we have a multivariate Normal, X and ε are independent.
- X is not a "cause" of Y, yet (1) still holds.
- X and Y have common "causes" Z_1 and Z_2, causing them to be correlated.
- Nonetheless, it is tempted to interpret (1) as a "response schedule".