

Econometrics 2 –Part 1

Arieda Muço

Central European University

Winter 2023

Linear Regression

- The conditional expectation function (CEF)
- CEF and regression
- Regression and causality
- The conditional independence assumption (CIA)
- Omitted variables bias
- Application: effect of computer use on wages

Statistical Model

1. **Step** Describe the *statistical* relationship between X_i and Y_i
2. **Step** *Econometric* interpretation of the relationship, causality

X_i, Y_i ... random variables

x_i, y_i $i = 1, \dots, N$ sample of random draws

Y_i dependent variable

X_i vector of covariates ($K \times 1$)

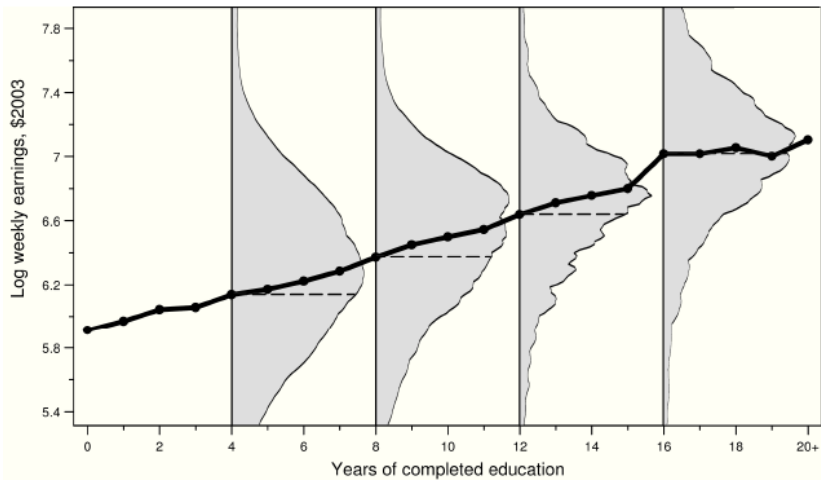
Conditional Expectation Function

$E[Y_i|X_i]$ expectation of Y_i holding X_i fixed

X_i random variable

$E[Y_i|X_i]$ random function

Example: conditional expectation of earnings given the level of schooling



CEF decomposition property

Theorem

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- (i) ϵ_i is mean independent of X_i , $E[\epsilon_i|X_i] = 0$
- (ii) ϵ_i is uncorrelated with any function of X_i

The CEF breaks Y_i into a part that is related to X_i and one that is *orthogonal* to X_i .

Regression Function

Best fitting line generated by minimizing the expected square errors

$$\underset{\beta}{\operatorname{argmin}} E [(Y_i - X_i' \beta)^2]$$

where β is a $(K \times 1)$ coefficient vector.

First order condition

$$\begin{aligned} E [X_i(Y_i - X_i' \beta)] &= 0 \\ \beta &= E [X_i X_i']^{-1} E [X_i Y_i] \end{aligned}$$

By construction the *residual* $e_i = Y_i - X_i' \beta$ is uncorrelated with X_i

Bivariate regression function

Best fitting line generated by minimizing the expected square errors

$$\underset{\alpha, \beta}{\operatorname{argmin}} E [(Y_i - \alpha - \beta X_i)^2]$$

First order condition gives

$$\begin{aligned}\beta &= \frac{\operatorname{Cov}(Y_i, X_i)}{\operatorname{Var}(X_i)} \\ \alpha &= E[Y_i] - \beta E[X_i]\end{aligned}$$

Estimator

$$\begin{aligned}\hat{\beta} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ \hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x}\end{aligned}$$

Multivariate regression function

$$Y_i = \beta_1 + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + e_i$$

Regression Anatomy Formula

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ik})}{Var(\tilde{x}_{ik})}$$

where \tilde{x}_{ik} is the residual from a regression of x_{ki} on all other X variables.

Proof

$$\begin{aligned} Cov(Y_i, \tilde{x}_{ik}) &= Cov(\beta_1 + \dots + \beta_K x_{Ki} + e_i, \tilde{x}_{ik}) \\ &= Cov(\beta_k x_{ki}, \tilde{x}_{ik}) = \beta_k Var(\tilde{x}_{ik}) \end{aligned}$$

Regression Function and CEF

Theorem 1 (linear CEF theorem):

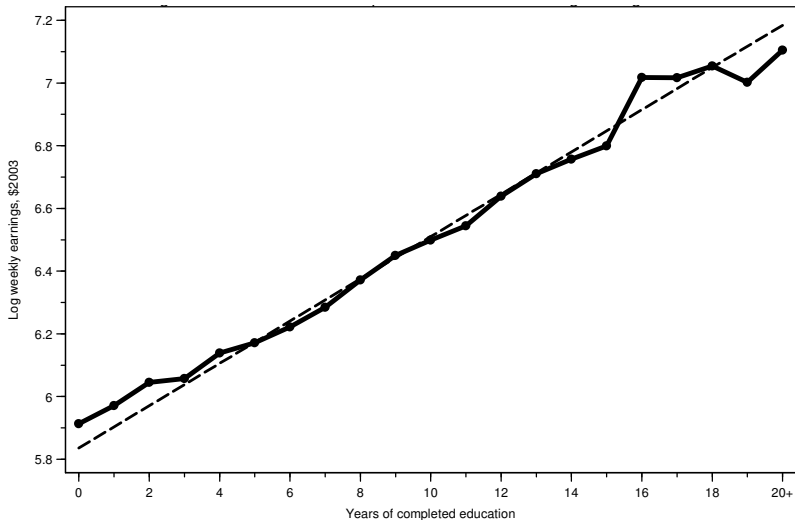
Suppose the CEF is linear, then the regression function is the CEF

Theorem 2 (regression CEF theorem):

$X_i'\beta$ provides the MMSE (minimum mean square error) linear approximation to $E[Y_i|X_i]$:

$$\beta = \underset{b}{\operatorname{argmin}} E \{ (E[Y_i|X_i] - X_i'b)^2 \}$$

Even if the CEF is not linear, regression provides the best linear approximation.



Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

Saturated Models

- Regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables.

Example 1:

y_i hourly wages

s_i schooling, $s_i = 0, 1, \dots, \tau$

- Saturated regression

$$\begin{aligned}y_i &= \alpha + \beta_1 d_{1i} + \dots + \beta_\tau d_{\tau i} + \epsilon_i \\d_{ij} &= 1 [s_i = j]\end{aligned}$$

Saturated Models

- Indicator variables are indicating each schooling level
- β_j j-th level of schooling effect

$$\begin{aligned}\beta_j &= E[Y_i | s_i = j] - E[Y_i | s_i = 0] \\ \alpha &= E[Y_i | s_i = 0], s_i = 0 \text{ (reference group)}\end{aligned}$$

- One parameter for any possible j
- Saturated models perfectly fit the CEF
(CEF is linear function of the set of indicator variables)

Saturated Models

Example 2:

- Two explanatory variables: College graduate and Sex

$$y_i = \alpha + \beta_1 coll + \beta_2 fem + \beta_3 coll \times female + \epsilon_i$$

- β_1 and β_2 main effects; β_3 interaction term.
- Other Parametrizations:

$$y_i = \tilde{\beta}_1 male_coll + \tilde{\beta}_2 male_nocoll + \tilde{\beta}_3 female_coll + \tilde{\beta}_4 female_nocoll + \epsilon_i$$

- CEF has four possible values: x_{1i} college and x_{2i} female

$$E[Y_i | x_{1i} = 0, x_{2i} = 0] = \alpha$$

$$E[Y_i | x_{1i} = 1, x_{2i} = 0] = \alpha + \beta_1$$

$$E[Y_i | x_{1i} = 0, x_{2i} = 1] = \alpha + \beta_2$$

$$E[Y_i | x_{1i} = 1, x_{2i} = 1] = \alpha + \beta_1 + \beta_2 + \beta_3$$

Saturated Models

Example 3:

- Multivalued schooling and Sex

$$y_i = \alpha + \sum_{j=1}^{\tau} \beta_j d_{ij} + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ij} \times x_{2i}) + \epsilon_i$$

- CEF has $(\tau + 1) \times 2$ values
- Saturated Model exactly fits CEF
- Increasingly restrictive model: omitting some interactions
e.g: Return to schooling is same for men and women
- But often models do not make sense! e.g:

$$y_i = \alpha + x_{2i} + \sum \delta_j (d_{ji} \times x_{2i}) + \epsilon_i$$

different returns to schooling levels *only* for women

Causality

- Regression causal \Leftrightarrow CEF causal
- Causal Effect: difference in average *potential outcome*
- The CEF is causal if it describes differences in the average potential outcomes for a fixed reference group
- Examples:
 - ▶ What would a given person earn if she obtained a different schooling level
 - ▶ What would people earn, on average, if we could change their schooling in a perfectly controlled environment
- Remember: in experiments causal variable is independent from potential outcomes

Conditional Independence Assumption (CIA)

Idea:

- Causal variable is independent of potential outcome conditional on observable variables X (ability, family, etc.)

Potential outcome framework

$$c_i \quad \text{College}$$
$$Y_i = \begin{cases} Y_{1i} & \text{if } c_i = 1 \\ Y_{0i} & \text{if } c_i = 0 \end{cases}$$

- Causal Effect $Y_{1i} - Y_{0i}$
- We want to estimate average $Y_{1i} - Y_{0i}$ for some group
- e.g. $E[Y_{1i} - Y_{0i} | c_i = 1]$ average causal effect for those who went to college.
- Remember:

$$\begin{aligned} E[Y_i | c_i = 1] - E[Y_i | c_i = 0] &= E[Y_{1i} - Y_{0i} | c_i = 1] \\ &+ \underbrace{E[Y_{0i} | c_i = 1] - E[Y_{0i} | c_i = 0]}_{\text{selection bias}} \end{aligned}$$

The conditional independence assumption asserts that conditional on observable variables X_i selection bias disappears:

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp c_i | X_i$$

$$\underbrace{E(Y_i | X_i, c_i = 1) - E(Y_i | X_i, c_i = 0)}_{\text{observed difference} \mid X} = \underbrace{E(Y_{1i} - Y_{0i} | X_i)}_{\text{difference in pot outcome} \mid X}$$

Multivalued Case

Multivalued causal variable s_i : level of schooling

Potential earnings for individual i with s years of schooling

$$Y_{s_i} = f_i(s)$$

Conditional independence assumption:

$$Y_{s_i} \perp\!\!\!\perp s_i | X_i \quad \text{for all } s$$

s_i is *as good as randomly assigned* conditional on X_i

Causal interpretation:

$$E[Y_i | X_i, s_i = s] - E[Y_i | X_i, s_i = s - 1] = E[f_i(s) - f_i(s - 1) | X_i]$$

- In the potential outcome framework we get a causal effect for every value of X_i
- Average over X_i using iterated expectations

$$\begin{aligned} & E \{ E [Y_i | X_i, s_i = 12] - E [Y_i | X_i, s_i = 11] \} \\ &= E \{ E [f_i(12) - f_i(11) | X_i] \} = E [f_i(12) - f_i(11)] \end{aligned}$$

- We also get a separate causal effect for each pair of levels of s_i , e.g. (11, 12), (12, 13),
- Use regression to *summarize* the effects

Regression

- Assumption: $f_i(s)$ is linear in s and the same for all i
- We estimate a weighted average of individual specific $f_i(s) - f_i(s-1)$

$$f_i(s) = \alpha + \rho s + \eta_i$$

- η_i error component: unobserved factors determining potential earnings
- Plug in observed values

$$Y_i = \alpha + \rho s_i + \eta_i$$

- Problem: s_i may be correlated with potential outcomes $f_i(s)$, or η_i .

Conditional Independence

$$\eta_i = X_i' \gamma + v_i$$

where γ is a vector of population regression coefficients

$$E[\eta_i | X_i] = X_i' \gamma$$

important: the residual v_i is uncorrelated with X_i

Suppose CIA holds given X_i , we can write:

$$\begin{aligned} E[f_i(s) | X_i, s_i] &= E[f_i(s) | X_i] = \alpha + \rho s_i + E[\eta_i | X_i] \\ &= \alpha + \rho s_i + X_i' \gamma \end{aligned}$$

We get the linear causal model:

$$Y_i = \alpha + \rho s_i + X_i' \gamma + v_i$$

Linear causal model:

$$Y_i = \alpha + \rho s_i + X_i' \gamma + v_i$$

- v_i is not correlated with s_i , or X_i . ρ represents causal effect.
- Key assumption: the observable variables X_i are the only reason that η_i , and s_i are correlated (or $f_i(s)$, and s_i)

Omitted Variable Bias

OVB formula describes relationship between regression models with different sets of control variables.

- Short Regression:

$$Y_i = \tilde{\alpha} + \tilde{\rho}s_i + \eta_i$$

- Long Regression:

$$Y_i = \alpha + \rho s_i + A_i\gamma + e_i$$

where A_i represents ability, family background.

- If CIA applies given A_i then ρ is the coefficient of the linear causal model.

Omitted Variables Bias Formula

$$\hat{\rho} = \frac{Cov(Y_i, s_i)}{V(s_i)} = \rho + \gamma' \delta_{As}$$

where δ_{As} vector of coefficients from regressions of A_i on s_i

coef from short regression = coef from long regression *plus* effect of omitted variables * regression coef of omitted on included variables

$$\begin{aligned} \frac{Cov(Y_i, s_i)}{V(s_i)} &= \frac{Cov(\alpha + \rho s_i + A_i' \gamma + e_i, s_i)}{V(s_i)} \\ &= \frac{\rho V(s_i) + \gamma Cov(A_i, s_i)}{V(s_i)} \\ &= \rho + \gamma \frac{Cov(A_i, s_i)}{V(s_i)} \end{aligned}$$

Interpretation

- $\delta_{As} = 0$, if A_i and s_i are uncorrelated
- Consequences of omitting A_i

Sign of bias

	$Cov(A_i, s_i) > 0$	$Cov(A_i, s_i) < 0$
$\gamma > 0$	positive	negative
$\gamma < 0$	negative	positive

Table 3.2.1: Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupation dummies
	0.132 (0.007)	0.131 (0.007)	0.114 (0.007)	0.087 (0.009)	0.066 (0.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2434.

*Additional controls are mother's and father's years of schooling and dummy variables for race and Census region.

Bad Controls

- Control for covariates (in some instances) allows for a causal interpretation of the main relationship.
Should we include any controls? Assume CIA holds.
- Caution:
 - ▶ Including bad controls might lead to bias if they are affected by the treatment.
- Example:
 - ▶ Randomly assign college degree (causal effect of mean earnings)
 - ▶ Occupation correlated with earnings and education.
- Should we control for occupation? (white collar job)

Application: Returns to Computer Use

- Alan B. Krueger, “How Computers Have Changed the Wage Structure: Evidence from Microdata 1984 - 1989”, QJE 1993
- John E. DiNardo and Jorn-Steffen Pischke, “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?”, QJE 2004

Returns to Computer Use: Motivation

Facts:

- Returns to education rising over time.
- Potential explanation: Skill Biased Technological Change
- Empirical evidence: Returns to computer use

Seminal study Krueger (1993)

- Cross sectional data on wages, computer on the job, CPS 1984, 1989
- Causal effect of computer use on wages plus 15-20% increase

Problem:

- Is there any unobservable characteristic x^* correlated to computer use? So that all individuals with this characteristic would earn higher wages anyway?

TABLE I
PERCENT OF WORKERS IN VARIOUS CATEGORIES WHO DIRECTLY
USE A COMPUTER AT WORK

Group	1984	1989
All workers	24.6	37.4
<u>Gender</u>		
Men	21.2	32.3
Women	29.0	43.4
<u>Education</u>		
Less than high school	5.0	7.8
High school	19.3	29.3
Some college	30.6	45.3
College	41.6	58.2
Postcollege	42.8	59.7
<u>Race</u>		
White	25.3	38.5
Black	19.4	27.7
<u>Age</u>		
Age 18–24	19.7	29.4
Age 25–39	29.2	41.5
Age 40–54	23.6	39.1
Age 55–65	16.9	26.3
<u>Occupation</u>		
Blue-collar	7.1	11.6
White-collar	33.0	48.4
<u>Union status</u>		
Union member	20.2	32.5
Nonunion	28.0	41.1
<u>Hours</u>		
Part-time	23.7	36.3
Full-time	28.9	42.7
<u>Region</u>		
Northeast	25.5	38.0
Midwest	23.4	36.0
South	23.2	36.5
West	27.0	39.9

Source. Author's tabulations of the 1984 and 1989 October Current Population Surveys. The sample size is 61,712 for 1984 and 62,748 for 1989.

TABLE II
OLS REGRESSION ESTIMATES OF THE EFFECT OF COMPUTER USE ON PAY
 (DEPENDENT VARIABLE: \ln (HOURLY WAGE))

Independent variable	October 1984			October 1989		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	1.937 (0.005)	0.750 (0.023)	0.928 (0.026)	2.086 (0.006)	0.905 (0.024)	1.094 (0.026)
Uses computer at work (1 = yes)	0.276 (0.010)	0.170 (0.008)	0.140 (0.008)	0.325 (0.009)	0.188 (0.008)	0.162 (0.008)
Years of education	—	0.069 (0.001)	0.048 (0.002)	—	0.075 (0.002)	0.055 (0.002)
Experience	—	0.027 (0.001)	0.025 (0.001)	—	0.027 (0.001)	0.025 (0.001)
Experience-squared ÷ 100	—	-0.041 (0.002)	-0.040 (0.002)	—	-0.041 (0.002)	-0.040 (0.002)
Black (1 = yes)	—	-0.098 (0.013)	-0.066 (0.012)	—	-0.121 (0.013)	-0.092 (0.012)
Other race (1 = yes)	—	-0.105 (0.020)	-0.079 (0.019)	—	-0.029 (0.020)	-0.015 (0.020)
Part-time (1 = yes)	—	-0.256 (0.010)	-0.216 (0.010)	—	-0.221 (0.010)	-0.183 (0.010)
Lives in SMSA (1 = yes)	—	0.111 (0.007)	0.105 (0.007)	—	0.138 (0.007)	0.130 (0.007)
Veteran (1 = yes)	—	0.038 (0.011)	0.041 (0.011)	—	0.025 (0.012)	0.031 (0.011)
Female (1 = yes)	—	-0.162 (0.012)	-0.135 (0.012)	—	-0.172 (0.012)	-0.151 (0.012)
Married (1 = yes)	—	0.156 (0.011)	0.129 (0.011)	—	0.159 (0.011)	0.143 (0.011)
Married*Female	—	-0.168 (0.015)	-0.151 (0.015)	—	-0.141 (0.015)	-0.131 (0.015)
Union member (1 = yes)	—	0.181 (0.009)	0.194 (0.009)	—	0.182 (0.010)	0.189 (0.010)
8 Occupation dummies	No	No	Yes	No	No	Yes
R^2	0.051	0.446	0.491	0.082	0.451	0.486

Notes. Standard errors are shown in parentheses. Sample size is 13,335 for 1984 and 13,379 for 1989. Columns (2), (3), (5), and (6) also include three region dummy variables.

Robustness Checks

- Coefficient gets smaller when other variables added
- Computer use at home and computer use at work.
- Narrow occupations: Secretaries 16% of comp. users in 1984, and 77% in 1989.
- Test Scores, Parental background from High school and beyond survey.

TABLE III
THE RETURN TO VARIOUS USES OF COMPUTERS, OCTOBER 1989^a
(DEPENDENT VARIABLE: \ln (HOURLY WAGE))

Use of computer at work	Proportion	Coefficient (std. error)
Uses computer at work for any task ^b	0.398	0.145 (0.010)
<u>Specific Task^c</u>		
Word processing	0.165	0.017 (0.012)
Bookkeeping	0.100	-0.058 (0.013)
Computer-assisted design	0.039	0.026 (0.020)
Electronic mail	0.063	0.149 (0.016)
Inventory control	0.102	-0.056 (0.013)
Programming	0.077	0.052 (0.031)
Desktop publishing or newsletters	0.036	-0.047 (0.021)
Spread sheets	0.094	0.079 (0.015)
Sales	0.060	-0.002 (0.016)
Computer games	0.019	-0.109 (0.026)
R^2		0.495

a. The sample and other explanatory variables are the same as in column (6) of Table II.

b. The computer use dummy variable equals one if the worker uses computers for any of the ten enumerated tasks or for any other task.

c. The dummy variables for any specific computer task, and the dummy variable for any computer use, are not mutually exclusive.

TABLE IV
THE RETURN TO COMPUTER USE AT WORK, HOME, AND WORK AND HOME
(STANDARD ERRORS ARE SHOWN IN PARENTHESES.)

Type of computer use	October 1984 (1)	October 1989 (2)	Percent of sample, 1989 (3)
Uses computer at work	0.165 (0.009)	0.177 (0.009)	39.8
Uses computer at home	0.056 (0.021)	0.070 (0.019)	12.5
Uses computer at home and work	0.006 (0.029)	0.017 (0.023)	8.6
Sample size	13,335	13,379	

Notes. The table reports coefficients for three dummy variables estimated from log hourly wage regressions. The other explanatory variables in the regressions are education, experience and its square, two race dummies, three region dummies, dummy variables indicating part-time status, residence in an SMSA, veteran status, gender, marital status, union membership, and an interaction between marital status and gender. Covariates are the same as in columns (2) and (5) of Table II.

TABLE V
OLS WAGE REGRESSION ESTIMATES FOR SECRETARIES
(DEPENDENT VARIABLE: \ln (HOURLY WAGE))

Independent variable	October 1984 (1)	October 1989 (2)
Intercept	1.387 (0.019)	1.208 (0.180)
Uses computer at work (1 = yes)	0.059 (0.024)	0.093 (0.030)
Years of education	0.014 (0.008)	0.035 (0.008)
Experience	0.009 (0.003)	0.024 (0.004)
Experience-squared \div 100	-0.007 (0.008)	-0.047 (0.009)
Black (1 = yes)	-0.079 (0.012)	0.065 (0.053)
Other race (1 = yes)	-0.095 (0.080)	0.065 (0.074)
Part-time (1 = yes)	-0.321 (0.031)	-0.160 (0.034)
Lives in SMSA (1 = yes)	0.159 (0.024)	0.152 (0.025)
Female (1 = yes)	0.090 (0.166)	0.146 (0.127)
Married (1 = yes)	0.422 (0.219)	-0.027 (0.027)
Married*Female	-0.387 (0.220)	—
Union member (1 = yes)	0.016 (0.040)	0.046 (0.046)
R^2	0.256	0.222

Notes. Standard errors are shown in parentheses. Sample size is 751 for 1984 and 618 for 1989. Regressions also include three region dummy variables. Mean (standard deviation) of the dependent variable for column (1) is 1.86 (0.36), and for column (2) is 2.08 (0.34).

TABLE VII
OLS REGRESSION ESTIMATES OF THE EFFECT OF COMPUTER USE ON PAY
(DEPENDENT VARIABLE: \ln (HOURLY WAGE))

Independent variable	October 1984			October 1989		
	(1)	(2)	(3)	(4)	(5)	(6)
Uses computer at work (1 = yes)	—	0.170 (0.008)	0.073 (0.048)	—	0.188 (0.008)	0.005 (0.043)
Computer use*Education	—	—	0.007 (0.003)	—	—	0.013 (0.003)
Years of education	0.076 (0.001)	0.069 (0.001)	0.067 (0.002)	0.086 (0.001)	0.075 (0.001)	0.071 (0.002)
Experience	0.027 (0.001)	0.027 (0.001)	0.027 (0.001)	0.027 (0.001)	0.027 (0.001)	0.027 (0.001)
Experience-squared \div 100	-0.042 (0.002)	-0.041 (0.002)	-0.042 (0.002)	-0.044 (0.002)	-0.041 (0.002)	-0.042 (0.002)
Black (1 = yes)	-0.106 (0.013)	-0.098 (0.013)	-0.099 (0.013)	-0.141 (0.013)	-0.121 (0.013)	-0.122 (0.013)
Other race (1 = yes)	-0.120 (0.020)	-0.105 (0.020)	-0.106 (0.020)	-0.037 (0.021)	-0.029 (0.020)	-0.032 (0.020)
Part-time (1 = yes)	-0.287 (0.010)	-0.256 (0.010)	-0.256 (0.010)	-0.261 (0.010)	-0.221 (0.010)	-0.221 (0.010)
Lives in SMSA (1 = yes)	0.123 (0.007)	0.111 (0.007)	0.111 (0.007)	0.148 (0.007)	0.138 (0.007)	0.138 (0.007)
Veteran (1 = yes)	0.043 (0.011)	0.038 (0.011)	0.039 (0.011)	0.027 (0.012)	0.025 (0.012)	0.029 (0.012)
Female (1 = yes)	-0.140 (0.012)	-0.162 (0.012)	-0.160 (0.012)	-0.142 (0.012)	-0.172 (0.012)	-0.168 (0.012)
Married (1 = yes)	0.162 (0.011)	0.156 (0.011)	0.156 (0.011)	0.169 (0.011)	0.159 (0.011)	0.158 (0.011)
Married*Female	-0.171 (0.015)	-0.168 (0.015)	-0.168 (0.015)	-0.146 (0.015)	-0.141 (0.015)	-0.139 (0.015)
Union member (1 = yes)	0.167 (0.009)	0.181 (0.009)	0.181 (0.009)	0.164 (0.010)	0.182 (0.010)	0.182 (0.010)
R^2	0.429	0.446	0.446	0.428	0.451	0.452
Mean-squared error	0.168	0.163	0.163	0.176	0.169	0.169

Notes. Standard errors are shown in parentheses. Sample size is 13,335 for 1984 and 13,379 for 1989. Regressions also include three region dummy variables and an intercept.

Returns to Computer Use Revisited

DiNardo and Pischke (2004):

- German data, also cross section, but includes multitude of workplace tools i.e: office tools, hand tools etc.
- Qualification and Career Survey: 1985/86, 1991/92
- Check comparability of Germany vs. US
- Are the coefficients similar? External validity.
- What is the impact of other workplace tools on wages?

TABLE I
PERCENT OF WORKERS IN VARIOUS CATEGORIES WHO USE DIFFERENT TOOLS
ON THEIR JOB

Group	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
Percentage that are computer users						
All workers	25.1	37.4	46.6	8.5	18.5	35.3
Men	21.6	32.2	41.1	7.9	18.5	36.4
Women	29.6	43.8	53.2	9.7	18.5	33.5
Less than high school	5.1	7.7	10.4	3.2	4.3	9.9
High school	19.2	28.4	34.6	8.5	18.3	32.7
Some college	30.6	45.0	53.1	8.5	24.8	48.4
College	42.4	58.8	70.2	13.4	30.5	61.6
Age 18–24	20.5	29.6	34.3	10.1	13.8	27.8
Age 25–39	29.6	41.4	49.8	9.6	21.6	39.9
Age 40–54	23.9	38.9	50.0	6.6	17.2	35.9
Age 55–64	17.7	27.0	37.3	5.9	13.5	23.7
Blue-collar	7.1	11.2	56.6	1.2	3.5	10.7
White-collar	39.7	56.6	67.6	12.8	28.9	50.2
Part-time	14.8	24.4	29.3	6.4	14.7	26.5
Full-time	29.3	42.3	51.0	8.7	19.1	37.0
Percentage of all workers who use a specific tool						
Computer	25.1	37.4	46.6	8.5	18.5	35.3
Calculator				19.6	35.7	44.2
Telephone				41.8	43.7	58.4
Pen/pencil				54.9	53.4	65.6
Work while sitting ^a				30.8	19.3	—
Hand tool (e.g., hammer)				29.4	32.9	30.5
Number of obs.	61,704	62,748	59,852	19,427	22,353	20,042

a. Variable definition differs in 1979 and 1985–1986. In 1979 it refers to “Never or rarely standing,” and in 1985–1986 it refers to “Often or almost always sitting.”

Columns 1 to 3 are from Table 3 in Autor, Katz, and Krueger [1996] and come from the October *Current Population Survey*. German data are from the *Qualification and Career Survey*.

TABLE II
OLS REGRESSIONS FOR THE EFFECT OF COMPUTER USE ON PAY
DEPENDENT VARIABLE: LOG HOURLY WAGE
(STANDARD ERRORS IN PARENTHESES)

Independent variable	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
Computer	0.171 (0.008)	0.188 (0.008)	0.204 (0.008)	0.112 (0.010)	0.157 (0.007)	0.171 (0.006)
Years of schooling	0.068 (0.001)	0.075 (0.002)	0.081 (0.002)	0.073 (0.001)	0.063 (0.001)	0.072 (0.001)
Experience	0.028 (0.001)	0.028 (0.001)	0.026 (0.001)	0.030 (0.001)	0.035 (0.001)	0.030 (0.001)
Experience ² /100	-0.043 (0.002)	-0.043 (0.002)	-0.041 (0.003)	-0.052 (0.002)	-0.058 (0.002)	-0.046 (0.002)
R ²	0.444	0.448	0.424	0.267	0.280	0.336
Number of obs.	13,335	13,379	13,305	19,427	22,353	20,042

Columns 1 to 3 are from Table 4 in Autor, Katz, and Krueger [1996]. Data for columns 1 to 3 are from the October *Current Population Survey*; data for columns 4 to 6 are from the *Qualification and Career Survey*. All models also include an intercept, a dummy for part-time, large city/MSA status, female, married, female*married. Regressions for the United States in columns 1 to 3 also include dummies for black, other race, veteran status, union membership, and three regions. Regressions for Germany in columns 4 to 6 also include a dummy for civil servants (*Beamter*).

Omitted Variable Bias

- Table III
 - ▶ Separate regressions for each tool

$$y_i = \beta_0 + \beta_1 tool + \beta_2 educ + \dots + u$$

- Table III-b
 - ▶ Enter tools together to check for correlation in use

$$y_i = \beta_0 + \beta_1 comp + \beta_2 calc + \beta_3 telep + \dots + u$$

TABLE III
OLS REGRESSION FOR THE EFFECT OF DIFFERENT TOOLS ON PAY
DEPENDENT VARIABLE: LOG HOURLY WAGE
(STANDARD ERRORS IN PARENTHESES)

Independent variable	Germany 1979	Germany 1985–86	Germany 1991–92	Germany 1979	Germany 1979	Germany 1985–1986	Germany 1991–1992
Occupation indicators	No	No	No	501	501	742	1071
Grades and father's Occupation*	No	No	No	No	Yes	No	No
Tools entered separately							
Computer	0.112 (0.010)	0.157 (0.007)	0.171 (0.006)	0.025 (0.011)	0.022 (0.011)	0.076 (0.008)	0.083 (0.007)
Calculator	0.087 (0.007)	0.128 (0.006)	0.129 (0.006)	0.027 (0.008)	0.025 (0.008)	0.061 (0.007)	0.054 (0.006)
Telephone	0.131 (0.006)	0.114 (0.006)	0.136 (0.006)	0.060 (0.007)	0.057 (0.007)	0.059 (0.007)	0.072 (0.007)
Pen/pencil	0.123 (0.006)	0.112 (0.006)	0.127 (0.006)	0.055 (0.007)	0.052 (0.007)	0.055 (0.007)	0.050 (0.007)
Work while sitting	0.106 (0.006)	0.101 (0.007)	—	0.042 (0.008)	0.041 (0.008)	0.036 (0.008)	—
Hand tool (e.g., hammer)	-0.117 (0.007)	-0.086 (0.006)	-0.091 (0.006)	-0.048 (0.009)	-0.045 (0.009)	-0.020 (0.008)	-0.020 (0.008)

	Tools entered together						
Computer	0.066 (0.010)	0.105 (0.008)	0.126 (0.007)	0.027 (0.011)	0.024 (0.011)	0.067 (0.008)	0.069 (0.007)
Calculator	0.017 (0.008)	0.053 (0.007)	0.044 (0.007)	0.015 (0.008)	0.014 (0.008)	0.032 (0.008)	0.022 (0.007)
Telephone	0.072 (0.007)	0.043 (0.008)	0.045 (0.008)	0.043 (0.008)	0.041 (0.008)	0.035 (0.008)	0.048 (0.008)
Pen/pencil	0.062 (0.007)	0.031 (0.008)	0.035 (0.008)	0.040 (0.008)	0.038 (0.008)	0.024 (0.008)	0.007 (0.008)
Work while sitting	0.058 (0.007)	0.050 (0.007)	—	0.036 (0.008)	0.035 (0.008)	0.032 (0.008)	—

a. Two variables for self-reported grades in math and German and eleven dummy variables for father's education.

Data are from the *Qualification and Career Survey*. All regressions also include an intercept, years of schooling, experience and experience squared, dummies for part-time, city, female, married, married*female, and for civil servants (*Beamter*).

Selection Problems

- What would be the ideal experiment?
- Randomly assign computers to workers?
- How are *computer skills* awarded in the labor market?
- What about the *skill* of using a pencil or a chair?
- Results point to substantial selection in who uses office tools

Bad Controls

- Compare mean earnings for white color workers with or without college.

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$

$$W_i = C_i W_{1i} + (1 - C_i) W_{0i}$$

$$E[Y_i | c_i = 1] - E[Y_i | c_i = 0] = E[Y_{1i} - Y_{0i}]$$

$$E[W_i | c_i = 1] - E[W_i | c_i = 0] = E[W_{1i} - W_{0i}]$$