Multiple regression

Data source: Current Population Survey 2006.

```
wage         average hourly earnings (in $)
educ         years of education
exper        years potential experience
female       1=female, 0=male
nonwhite     =1 if nonwhite
services     =1 if in services industry
profocc      =1 if in professional occupation
union        =1 if respondent is union member
```

Summary statistics of the variables

```
. summarize wage educ exper female nonwhite services profocc union

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |      2000    18.34701    11.49495         .7   82.42857
        educ |      2000      13.633      2.0877          9         18
       exper |      2000     20.5885    12.75769          0         67
      female |      2000       .5165    .4998527          0          1
    nonwhite |      2000        .149    .3561775          0          1
    services |      2000        .152    .3591107          0          1
     profocc |      2000        .212    .4088271          0          1
       union |      2000        .139     .346033          0          1
```

Regression of log(wage) on education (with increasing number of other controls)

$$\widehat{\log(wage)} = \underset{(.08)}{1.19} + \underset{(.006)}{.114}\,educ \qquad R^2 = .17,\ n = 2000$$

$$\widehat{\log(wage)} = \underset{(.08)}{0.97} + \underset{(.005)}{.114}\,educ + \underset{(.0009)}{.011}\,exp \qquad R^2 = .22,\ n = 2000$$

$$\widehat{\log(wage)} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,educ + \underset{(.0009)}{.011}\,exp - \underset{(.02)}{.25}\,female \qquad R^2 = .27,\ n = 2000$$

$$\widehat{\log(wage)} = \underset{(.08)}{1.21} + \underset{(.005)}{.109}\,educ + \underset{(.0009)}{.010}\,exp - \underset{(.02)}{.24}\,female - \underset{(.03)}{.23}\,services \qquad R^2 = .29,\ n = 2000$$

```
. generate lwage=log(wage)
. regress lwage educ exper female services

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  4,  1995) =  203.57
       Model |  194.949697     4  48.7374243           Prob > F      =  0.0000
    Residual |   477.62717  1995  .239412115           R-squared     =  0.2899
-------------+------------------------------           Adj R-squared =  0.2884
       Total |  672.576867  1999  .336456662           Root MSE      =   .4893

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1086052    .005367    20.24   0.000     .0980798    .1191306
       exper |   .0103387   .0008616    12.00   0.000     .0086489    .0120284
      female |  -.2437118   .0219738   -11.09   0.000    -.2868059   -.2006178
    services |  -.2272342   .0313323    -7.25   0.000    -.2886816   -.1657868
       _cons |    1.20755   .0776504    15.55   0.000     1.055266    1.359834
------------------------------------------------------------------------------
```

$$\widehat{\log(wage)} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,educ + \underset{(.0009)}{.011}\,exp - \underset{(.02)}{.25}\,female \qquad R^2 = .27,\ n = 2000$$

Adding/omitting an irrelevant variable:

$$\widehat{\log(wage)} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,educ + \underset{(.0009)}{.011}\,exp - \underset{(.02)}{.25}\,female - \underset{(.031)}{.037}\,nonwhite \qquad R^2 = .27,\ n = 2000$$

Adding/omitting an important variable not correlated with the other independent variables:

$$\widehat{\log(wage)} = \underset{(.08)}{1.28} + \underset{(.006)}{.117}\,educ - \underset{(.02)}{.25}\,female \qquad R^2 = .21,\ n = 2000$$

Adding/omitting an important variable correlated with the other independent variables:
Omitted variable bias

$$\widehat{\log(wage)} = \underset{(.08)}{1.17} + \underset{(.005)}{.106}\,educ + \underset{(.0009)}{.011}\,exp - \underset{(.02)}{.26}\,female + \underset{(.03)}{.012}\,profocc \qquad R^2 = .28,\ n = 2000$$

$$\widehat{\log(wage)} = \underset{(.03)}{2.57} + \quad + \underset{(.0009)}{.011}\,exp - \underset{(.02)}{.26}\,female + \underset{(.03)}{.358}\,profocc \qquad R^2 = .16,\ n = 2000$$

```
. correlate lwage educ exp female profocc nonwhite
(obs=2000)

             |    lwage     educ    exper   female  profocc nonwhite
-------------+------------------------------------------------------
       lwage |   1.0000
        educ |   0.4097   1.0000
       exper |   0.2358   0.0010   1.0000
      female |  -0.1935   0.0489   0.0210   1.0000
     profocc |   0.2181   0.4276  -0.0383   0.1077   1.0000
    nonwhite |  -0.0379  -0.0051  -0.0200   0.0368  -0.0143   1.0000
```

Interpretation

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ $\qquad\qquad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

β_1 measures the effect on y of a change in x_1 by 1 (unit), holding other factors fixed (both x_2 and u)
β_1 measures the effect on $E(y)$ of a change in x_1 by 1 (unit), holding x_2 fixed.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

$\hat{\beta}_1$ measures the effect on the predicted $\hat{y}$ of a change in x by 1 (unit), holding x_2 fixed.
"Holding experience and gender fixed, a one year increase in education is associated with a 11.7% increase in predicted wage"