Lecture 3: Multivariate Regression

Homework review

- Question C2.4 ask you to estimate a simple bivariate regression using IQ to predict wages.
- In Stata this looks like

```
. reg wage IQ
```

not

```
. reg IQ wage
```

- What does the latter command give you?

Homework review

```
. reg wage IQ

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =   98.55
       Model |  14589782.6     1  14589782.6           Prob > F      =  0.0000
    Residual |   138126386   933  148045.429           R-squared     =  0.0955
-------------+------------------------------           Adj R-squared =  0.0946
       Total |   152716168   934  163507.675           Root MSE      =  384.77

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   8.303064   .8363951     9.93   0.000     6.661631    9.944498
       _cons |   116.9916   85.64153     1.37   0.172    -51.08078    285.0639
------------------------------------------------------------------------------
```

- What is the predicted increase in monthly salary for a 15 point increase in IQ?
- Common mistake: 8.3*15 + 117
  - Why is this wrong?
- What is the predicted monthly salary for IQs of 100, 115, 145?

Explaining State Homicide Rates, cont.

- Two weeks ago, we modeled state homicide rates as being dependent on one variable: poverty. In reality, we know that state homicide rates depend on numerous variables.
- Our estimation of homicide rates using multiple regression will look something like this: $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$
- This allows us to estimate the “effect” of any one factor while holding “all else constant.”

Explaining State Homicide Rates, cont.

The "true" model:

$$Y_i = \mu_0 + \mu_1 E_{i1} + \mu_2 E_{i2} + \ldots + \mu_p E_{ip} + R_i$$

$$= \mu_0 + \sum_{j=1}^{p} \mu_j E_{ij} + R_i$$

Our estimation model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$$

Explaining State Homicide Rates, cont.

- Usually, the independent variables in our estimation model are some subset of the "true" model.
- We can rewrite the "true" model in terms of *k* observed and *p-k* unobserved variables:

$$Y_i = \mu_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \sum_{j=k+1}^{p} \mu_j E_{ij} + R_i$$

Explaining State Homicide Rates, cont.

- Re-arranging the "true" equation:

$$\sum_{j=1}^{k} \beta_j X_{ij} = (Y_i - \mu_0) - \sum_{j=k+1}^{p} \mu_j E_{ij} - R_i$$

- Re-arranging the estimation equation:

$$\varepsilon_i = Y_i - \beta_0 - \sum_{j=1}^{k} \beta_j X_{ij}$$

- And substituting:

$$\varepsilon_i = Y_i - \beta_0 - Y_i + \mu_0 + \sum_{j=k+1}^{p} \mu_j E_{ij} + R_i$$

$$= (\mu_0 - \beta_0) + \sum_{j=k+1}^{p} \mu_j E_{ij} + R_i$$

Explaining State Homicide Rates, cont.

- This means that the error term in a regression reflects both the random component in the dependent variable, and the impact of all excluded variables.
- Variables besides poverty thought to influence homicide rates:
  - Region, high school graduation, incarceration, unemployment, gun ownership, female headed households, population heterogeneity, income, welfare, law enforcement officers, IQ, smokers, other crime

Explaining State Homicide Rates, example

- Recall, in a bivariate regression, we found the following:

$$E(\hom rate_i) = -.973 + .475\, poverty_i + u_i$$

- Download multivariate homicide rate data "murder_multi.dta" from www.public.asu.edu/~gasweete/crj604/data/
- Adding imprisonment rate and rate of female-headed households to the model yields the following:

$$E(\hom rate_i) = -7.34 - .005\, poverty_i + .0077\, prison_i + .89\, femhh_i + u_i$$

Explaining State Homicide Rates, example

- Add imprisonment rate and rate of female-headed households to the regression model predicting homicide rates.
- You should get a model like this:

$$E(\text{hom}\,rate_i) = -7.34 - .005\,poverty_i + .0077\,prison_i + .89\,femhh_i + u_i$$

- What happened to the relationship between poverty and homicide? Why?
- What does it mean that our intercept is now -7.34?

Explaining State Homicide Rates, example

$$E(\hom rate_i) = -7.34 - .005\, poverty_i + .0077\, prison_i + .89\, femhh_i + u_i$$

- Of the three predictors in our model, which is the “strongest”?
- Poverty is no longer statistically significant. How precise is our estimate of the poverty effect? Hint: what is the 95% confidence interval?
  - Does this interval contain large effects. Another hint: what is the 95% confidence interval for the standardized coefficient?

Explaining State Homicide Rates, example

- In the bivariate regression, imprisonment rates and rates of female-headed households were in the error term, and assumed to be uncorrelated with poverty rates.
- This assumption was false. In fact, explicitly controlling for just these two variables reduces the estimate for the effect of poverty on homicide rates from .475 to -.005

Explaining State Homicide Rates, example

$$E(\hom rate_i) = -7.34 - .005\, poverty_i + .0077\, prison_i + .89\, femhh_i + u_i$$

- It's important to know how to interpret the regression results.
- -7.34 is the expected homicide rate if poverty rates, imprisonment rates, and female-headed household rates were zero. This is never the case, so it's not a meaningful estimate.
- .0077 is the effect of a 1 point increase in the imprisonment rate on the homicide rate, *holding poverty and femhh constant.*
- .89 is the effect of a 1 point increase in the female-headed household rate on the homicide rate, *holding poverty and prison constant.*
- See Wooldridge pp. 78-9 (partialling out)

Explaining State Homicide Rates, example

$$E(\hom rate_i) = -7.34 - .005\, poverty_i + .0077\, prison_i + .89\, femhh_i + u_i$$

- Is the effect of female-headed households 115 times bigger than the effect of the imprisonment rate?
- *prison*: mean=404, s.d.=141
- *femhh*: mean=10.2, s.d.=1.4
- Because the standard deviation of *prison* is 100 times larger than *femhh*, it’s not easy to directly compare the two estimates, unless we calculate standardized effects:
  - *prison*: .422, *femhh*: .499

Explaining State Homicide Rates, example

$$E(\hom rate_i) = -7.34 - .005\,poverty_i + .0077\,prison_i + .89\,femhh_i + u_i$$

- The fitted value (or predicted value) for each state is the expected homicide rate given the poverty, imprisonment and female-headed household rate.
- For Arizona:

$$E(\hom rate_i) = -7.34 - .005*15.2 + .0077*529 + .89*10.06$$

$$= -7.34 - .076 + 4.07 + 8.95$$

$$= 5.60$$

Explaining State Homicide Rates, example

$$E(\hom rate_i) = -7.34 - .005\,poverty_i + .0077\,prison_i + .89\,femhh_i + u_i$$

- The actual homicide rate in Arizona was 7.5, so the residual is 1.9

$$u_i = y_i - \hat{y}_i = 7.5 - 5.6 = 1.9$$

- That's just one of 50 residuals. The sum of all residuals is zero.
- The sum of the squares of all residuals is as small as possible. That's how the estimates are chosen

Explaining State Homicide Rates, example

- Rather than calculating the predicted values and residuals “by hand”, you can have Stata do it:
- For predicted values, after your regression model (“homhat” is the name of the new variable. It can be anything you want to call it.):

```
. predict homhat
(option xb assumed; fitted values)
```

- For residuals (again, “resid” can be anything):

```
. predict resid, r
```

Explaining State Homicide Rates, example

- You can also estimate predicted values for hypothetical cases.
- For example, if we wanted to look at the "average state":

Explaining State Homicide Rates, example

```
. summ poverty prison fem

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     poverty |        50       12.09        3.01        5.6       20.1
      prison |        50      405.34    141.3413        141        835
      fem_hh |        50    10.16597    1.445036   7.533991   14.21424

. reg homrate poverty prison fem_hh

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) =   30.34
       Model |  216.069398     3  72.0231325           Prob > F      =  0.0000
    Residual |  109.215601    46   2.3742522           R-squared     =  0.6642
-------------+------------------------------           Adj R-squared =  0.6423
       Total |  325.284999    49  6.63846936           Root MSE      =  1.5409

------------------------------------------------------------------------------
     homrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |  -.0048145   .1003278    -0.05   0.962    -.2067639    .1971349
      prison |   .0077024    .002054     3.75   0.000      .003568    .0118368
      fem_hh |    .889994   .2051466     4.34   0.000     .4770552    1.302933
       _cons |  -7.341531    1.59615    -4.60   0.000    -10.55441   -4.128649
------------------------------------------------------------------------------

. di _b[_cons]+_b[poverty]*12.09+_b[prison]*405.34+_b[fem_hh]*10.16597
4.7699969
```

Explaining State Homicide Rates, example

- We can also look at a more disadvantaged hypothetical state:

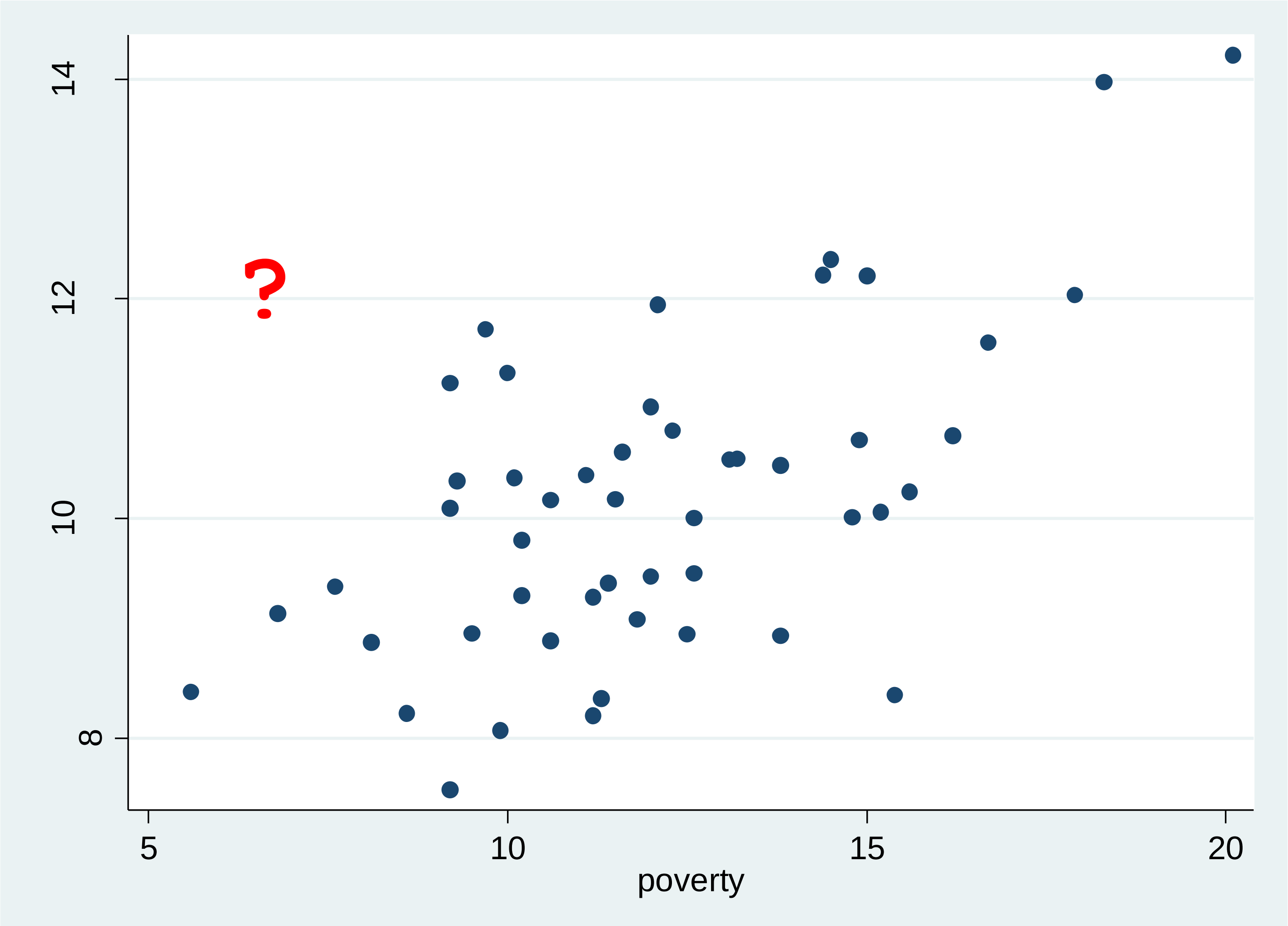
```
. di _b[_cons]+_b[poverty]*15+_b[prison]*600+_b[fem_hh]*12
7.8876081
```

- Or an unusual state, where poverty and imprisonment rates are low but female headed household rate is high:

```
. di _b[_cons]+_b[poverty]*7+_b[prison]*200+_b[fem_hh]*12
4.8451712
```

- Is this last prediction reasonable?

Explaining State Homicide Rates, example



R^2

- Estimating and interpreting R^2 remains the same in multivariate regression.

$$R^2 = \frac{SSE}{SST} = \frac{\sum \left(\hat{y}_i - \overline{y} \right)^2}{\sum \left(y_i - \overline{y} \right)^2}$$

- As more variables are included in the model, R^2 will either stay the same or increase.
- One danger is overfitting, where variables are included in the model that are "explaining" noise or random error in the dependent variable

R^2, example

```
. reg hom pov

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   21.36
       Model |  100.175656     1  100.175656           Prob > F      =  0.0000
    Residual |  225.109343    48  4.68977798           R-squared     =  0.3080
-------------+------------------------------           Adj R-squared =  0.2935
       Total |  325.284999    49  6.63846936           Root MSE      =  2.1656

------------------------------------------------------------------------------
     homrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |    .475025   .1027807     4.62   0.000     .2683706    .6816795
       _cons |  -.9730529   1.279803    -0.76   0.451     -3.54627    1.600164
------------------------------------------------------------------------------
```

R^2, example

```
. reg homrate pov IQ gdp leo welfare smokers income het gunowner fem_ unemp prison gradrate
    pop65

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F( 14,    35) =    7.57
       Model |  244.511494    14  17.4651067           Prob > F      =  0.0000
    Residual |  80.7735048    35  2.30781442           R-squared     =  0.7517
-------------+------------------------------           Adj R-squared =  0.6524
       Total |  325.284999    49  6.63846936           Root MSE      =  1.5191

------------------------------------------------------------------------------
     homrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |  -.1260969   .1570399    -0.80   0.427     -.444905    .1927111
          IQ |  -.2960415   .2012222    -1.47   0.150    -.7045442    .1124613
  gdp_percap |  -.0000843   .0000675    -1.25   0.220    -.0002214    .0000527
         leo |   .0078023   .0062672     1.24   0.221    -.0049209    .0205255
     welfare |   .0043498   .0046595     0.93   0.357    -.0051096    .0138091
     smokers |   .0731863   .0906833     0.81   0.425    -.1109106    .2572833
income_per~p |   .0000533   .0001005     0.53   0.599    -.0001508    .0002574
         het |   1.716118   3.287625     0.52   0.605    -4.958115    8.390351
    gunowner |   .0026661   .0301547     0.09   0.930    -.0585511    .0638834
      fem_hh |   .5857682   .2843154     2.06   0.047     .0085773    1.162959
```

R^2, example

- Our R^2 went up to .75! We can explain 75% of the variance in homicide rates, or can we? It could be that our high R^2 is due to overfitting.
- Solutions
  - If you have enough cases, split your sample and build your model on half the cases. Test it once on the remaining cases.
  - If you can't do that, avoid iterative or stepwise modeling as it produces biased estimates.
  - Pay more attention to adjusted R^2.

Adjusted R^2

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$$\tilde{R}^2 = 1 - \frac{SSR/(N-k)}{SST/(N-1)} = 1 - (1 - R^2)\frac{N-1}{N-k}$$

- Adjusted r-squared “penalizes” our estimate of explained variance by the number of parameters used.

F-test

$$F_{k-1,N-k} = \frac{SSE \big/ k-1}{SSR \big/ N-k} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1}$$

- The formula for the F-statistic remains the same in a multivariate context, we just have to adjust the degrees of freedom depending on how many parameters are in the model
- You can use the last expression above to calculate the F-statistic if Stata doesn't provide it, and all you have is R^2

F-test, cont.

- The F-test can be thought of as a formal test of the significance of R^2

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \exists\, j \in [1, k] : \beta_j \neq 0$$

- That last line reads: "There exists *j* in the set of values from 1 to k such that β_j does not equal zero." In other words, at least one variable is statistically significant.

Gauss-Markov Assumptions

1) **Linear in Parameters**:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

2) **Random Sampling**: we have a random sample from the population that follows the above model.
3) **No Perfect Collinearity:** None of the independent variables is a constant, and there is no exact linear relationship between independent variables.
4) **Zero Conditional Mean**: The error has zero expected value for each set of values of *k* independent variables: $E(\varepsilon_i) = 0$
5) **Unbiasedness of OLS**: The expected value of our beta estimates is equal to the population values (the true model).

Next time:

Homework: Problems 3.2, 3.4, C3.2, C3.4, C3.6

Read: Wooldridge Chapters 3 (again) & 4