Lecture 4: Multivariate Regression, Part 2

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).

Assumption MLR1: Linear in Parameters

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

- This assumption refers to the *population* or *true* model.
- Transformations of the *x* and *y* variables are allowed. But the dependent variable or its transformation must be a linear combination the *β* parameters.

Assumption MLR1: Common transformations

- Level-log: $y = \beta_0 + \beta_1 \log(x) + \varepsilon$
  - Interpretation: a one percent increase in x is associated with a (β_1/100) increase in y.

```
. gen lpov=log(poverty)

. reg homrate lpov

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   19.06
       Model |  92.4491443     1  92.4491443           Prob > F      =  0.0001
    Residual |  232.835854    48  4.85074697           R-squared     =  0.2842
-------------+------------------------------           Adj R-squared =  0.2693
       Total |  325.284999    49  6.63846936           Root MSE      =  2.2024

------------------------------------------------------------------------------
     homrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpov |   5.376775   1.231614     4.37   0.000     2.900448    7.853101
       _cons |  -8.463789   3.047317    -2.78   0.008    -14.59083   -2.336748
------------------------------------------------------------------------------
```

  - So a one percent increase in poverty results in an increase of .054 in the homicide rate
  - This type of relationship is not commonly used.

Assumption MLR1: Common transformations

- Log-level: $\log(y) = \beta_0 + \beta_1 x + \varepsilon$
  - Interpretation: a one unit increase in x is associated with a $(100 \cdot \beta_1)$ *percent* increase in y.

```
. gen lhom=log(homrate)

. reg lhom poverty

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   19.64
       Model |  5.48717761     1  5.48717761           Prob > F      =  0.0001
    Residual |  13.4093807    48  .279362099           R-squared     =  0.2904
-------------+------------------------------           Adj R-squared =  0.2756
       Total |  18.8965583    49  .385644048           Root MSE      =  .52855

------------------------------------------------------------------------------
        lhom |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |   .1111757   .0250853     4.43   0.000     .0607384    .1616131
       _cons |   .0500712   .3123567     0.16   0.873    -.5779641    .6781064
------------------------------------------------------------------------------
```

  - So a one unit increase in poverty (one percentage point) results in an 11.1% increase in homicide.

Assumption MLR1: Common transformations

- Log-log: $\log(y) = \beta_0 + \beta_1 \log(x) + \varepsilon$
  - Interpretation: a one percent increase in x is associated with a β_1 *percent* increase in y.

```
. reg lhom lpov

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   19.45
       Model |  5.44961433     1  5.44961433           Prob > F      =  0.0001
    Residual |   13.446944    48  .280144667           R-squared     =  0.2884
-------------+------------------------------           Adj R-squared =  0.2736
       Total |  18.8965583    49  .385644048           Root MSE      =  .52929

------------------------------------------------------------------------------
        lhom |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpov |   1.305429   .2959795     4.41   0.000     .7103225    1.900536
       _cons |  -1.818851    .732326    -2.48   0.017    -3.291291   -.3464107
------------------------------------------------------------------------------
```

  - So a one percent increase in poverty results in an 1.31% increase in homicide.
  - These three are explained on p. 46

Assumption MLR1: Common transformations

- Non-linear: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  - Interpretation: The relationship between x and y is not linear. It depends on levels of x.
  - A one unit change in x is associated with a $\beta_1 + 2 * \beta_2 * x$ change in y.

```
. reg homrate c.pov##c.pov

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   10.85
       Model |  102.719645     2  51.3598226           Prob > F      =  0.0001
    Residual |  222.565354    47  4.73543305           R-squared     =  0.3158
-------------+------------------------------           Adj R-squared =  0.2867
       Total |  325.284999    49  6.63846936           Root MSE      =  2.1761

------------------------------------------------------------------------------------
           homrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
           poverty |   .0020093   .6535657     0.00   0.998    -1.312795    1.316814
                   |
c.poverty#c.poverty|   .0186286   .0254157     0.73   0.467    -.0325012    .0697584
                   |
             _cons |   1.857401   4.070203     0.46   0.650    -6.330792    10.04559
------------------------------------------------------------------------------------
```

What the c.## is going on?

- You could create a new variable that is poverty squared and enter that into the regression model, but there are benefits to doing it the way I showed you on the previous slide.
- “c.” tells Stata that this is a continuous variable.
  - You can also tell Stata that you’re using a categorical variable with i. – and you can tell it which category to use as the base level with i2., i3., etc.
  - More info here: http://www.ats.ucla.edu/stat/stata/seminars/stata11/fv_seminar.htm

What the c.## is going on?

- ## tells Stata to control for the product of the variables on both sides as well as the variables themselves. In this case, since pov is on both sides, it controls for pov once, and pov squared.
  - Careful! Just one pound # between the variables would mean Stata would only control for the squared term – something we rarely if ever would want to do.
- The real benefit of telling Stata about squared terms or interaction terms is that Stata can then report accurate marginal effects using the "margins" command.

Assumption MLR1: Common transformations

- Non-linear: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  - Both the linear and squared poverty variables were not statistically significant in the previous regression, but they are jointly significant. (Look at the F-test).
  - When poverty goes from 5 to 6%, homicide goes up by (.002+2*5*.019)=.192
  - When poverty goes from 10 to 11%, homicide goes up by (.002+2*10*.019)=.382
  - From 19 to 20: .762
  - So this is telling us that the impact of poverty on homicide is worse when poverty is high.
  - You can also learn this using the margins command:

Assumption MLR1: Common transformations

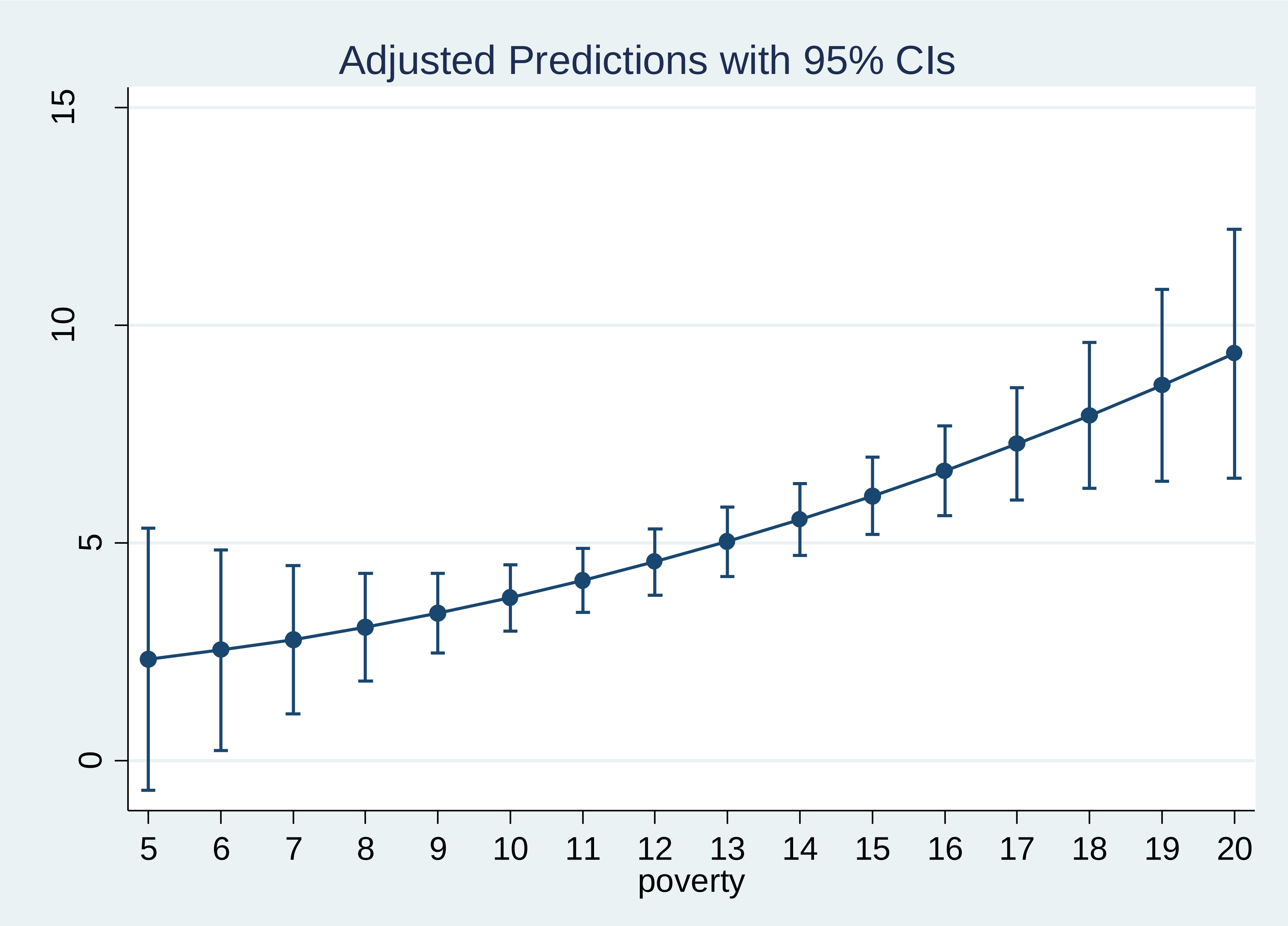
- `. margins, at(poverty=(5(1)20))`
  - This gives predicted values of the homicide rate for values of the poverty rate ranging from 5 to 20. If we follow this command with the "`marginsplot`" command, we'll see a nice graph depicting the non-linear relationship between poverty and homicide.
- `. margins, dydx(poverty) at(poverty=(5(1)20))`
  - This gives us the rate of change in homicide rate at different levels of poverty, showing that the change is greater at higher levels of poverty.

Assumption MLR1: Common transformations

- `. margins, at(poverty=(5(1)20))`

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf. Interval]	
_at						
1	2.333163	1.498207	1.56	0.126	-.6808403	5.347165
2	2.540086	1.143902	2.22	0.031	.2388523	4.84132
3	2.784267	.8477508	3.28	0.002	1.078813	4.489722
4	3.065705	.6156911	4.98	0.000	1.827095	4.304316
5	3.3844	.4573197	7.40	0.000	2.464392	4.304409
6	3.740353	.3792478	9.86	0.000	2.977405	4.503301
7	4.133562	.3655011	11.31	0.000	3.398269	4.868856
8	4.564029	.3799787	12.01	0.000	3.799611	5.328448
9	5.031753	.3970721	12.67	0.000	4.232947	5.830559
10	5.536734	.4127955	13.41	0.000	4.706297	6.367172
11	6.078972	.4416451	13.76	0.000	5.190497	6.967448
12	6.658468	.5094949	13.07	0.000	5.633496	7.683439
13	7.27522	.6383377	11.40	0.000	5.991051	8.55939
14	7.92923	.8352808	9.49	0.000	6.248862	9.609598
15	8.620497	1.097597	7.85	0.000	6.412417	10.82858
16	9.349021	1.42075	6.58	0.000	6.490841	12.2072

- Followed by `marginsplot:`



Adjusted Predictions with 95% CIs
0
5
10
15
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
poverty

Assumption MLR1: Common transformations

- Interaction: $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$
  - Interpretation: The relationship between x_1 and y depends on levels of x_2. And/or the relationship between x_2 and y depends on levels of x_1.
  - We'll cover interaction terms and other non-linear transformations later.
  - The best way to enter them into the regression model is to use the ## pattern as with squared terms so that the margins command will work properly and marginsplot will create cool graphs.

Assumption MLR2: Random Sampling

- We have a random sample of *n* observations from the population.
- Think about what your population is. If you modify the sample by dropping cases, you may no longer have a random sample from the original population, but you may have a random sample of another population.
  - Ex: relationship breakup and crime
- We'll deal with this issue in more detail later.

Assumption MLR3: No perfect collinearity

- None of the independent variables is a constant.
- There is no exact *linear* relationship among the independent variables.
- In practice, in either of these situations, one of the offending variables will be dropped from the analysis by Stata.
- High collinearity is *not* a violation of the regression assumptions, nor are nonlinear relationships among variables.

Assumption MLR3: No perfect collinearity, example

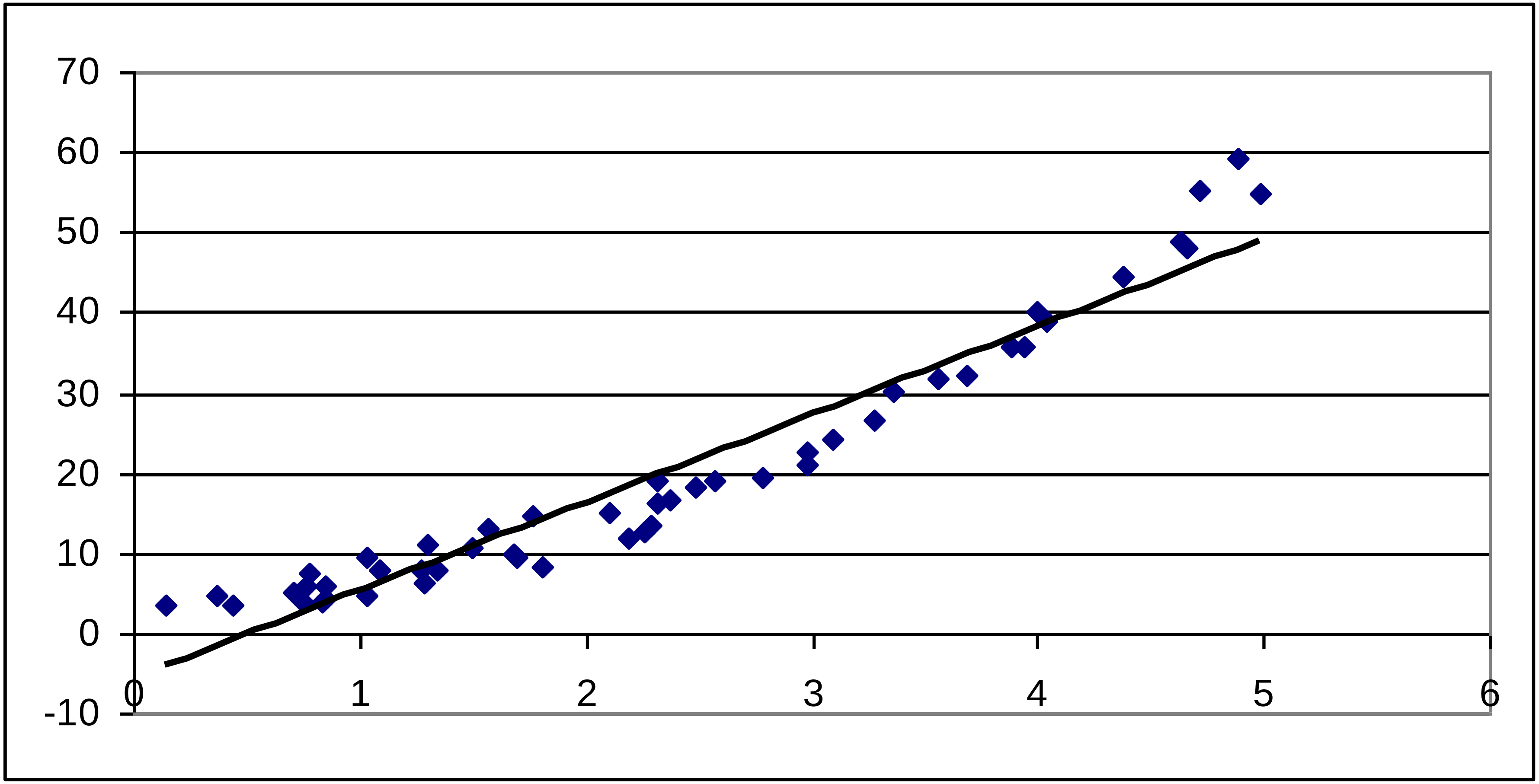
```
. reg dfreq7 male hisp white black first asian other age6 dropout6 dfreq6

      Source |       SS       df       MS              Number of obs =    6794
-------------+------------------------------           F(  9,  6784) =  108.26
       Model |  218609.566     9  24289.9518           Prob > F      =  0.0000
    Residual |  1522043.58  6784  224.357839           R-squared     =  0.1256
-------------+------------------------------           Adj R-squared =  0.1244
       Total |  1740653.14  6793  256.242182           Root MSE      =  14.979

------------------------------------------------------------------------------
      dfreq7 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.784668   .3663253     4.87   0.000     1.066556    2.502781
        hisp |   .4302673   .5786788     0.74   0.457    -.7041247    1.564659
       white |   1.225733   2.248439     0.55   0.586    -3.181912    5.633379
       black |   2.455362   2.267099     1.08   0.279    -1.988863    6.899587
       first |  (dropped)
       asian |  -.2740142   2.622909    -0.10   0.917    -5.415739     4.86771
       other |   1.309557    2.32149     0.56   0.573    -3.241293    5.860406
        age6 |  -.2785403   .1270742    -2.19   0.028    -.5276457    -.029435
    dropout6 |   .6016927    .485114     1.24   0.215    -.3492829    1.552668
      dfreq6 |   .3819413   .0128743    29.67   0.000     .3567037    .4071789
       _cons |   4.617339   3.365076     1.37   0.170    -1.979265    11.21394
------------------------------------------------------------------------------
```

Assumption MLR4: Zero Conditional Mean

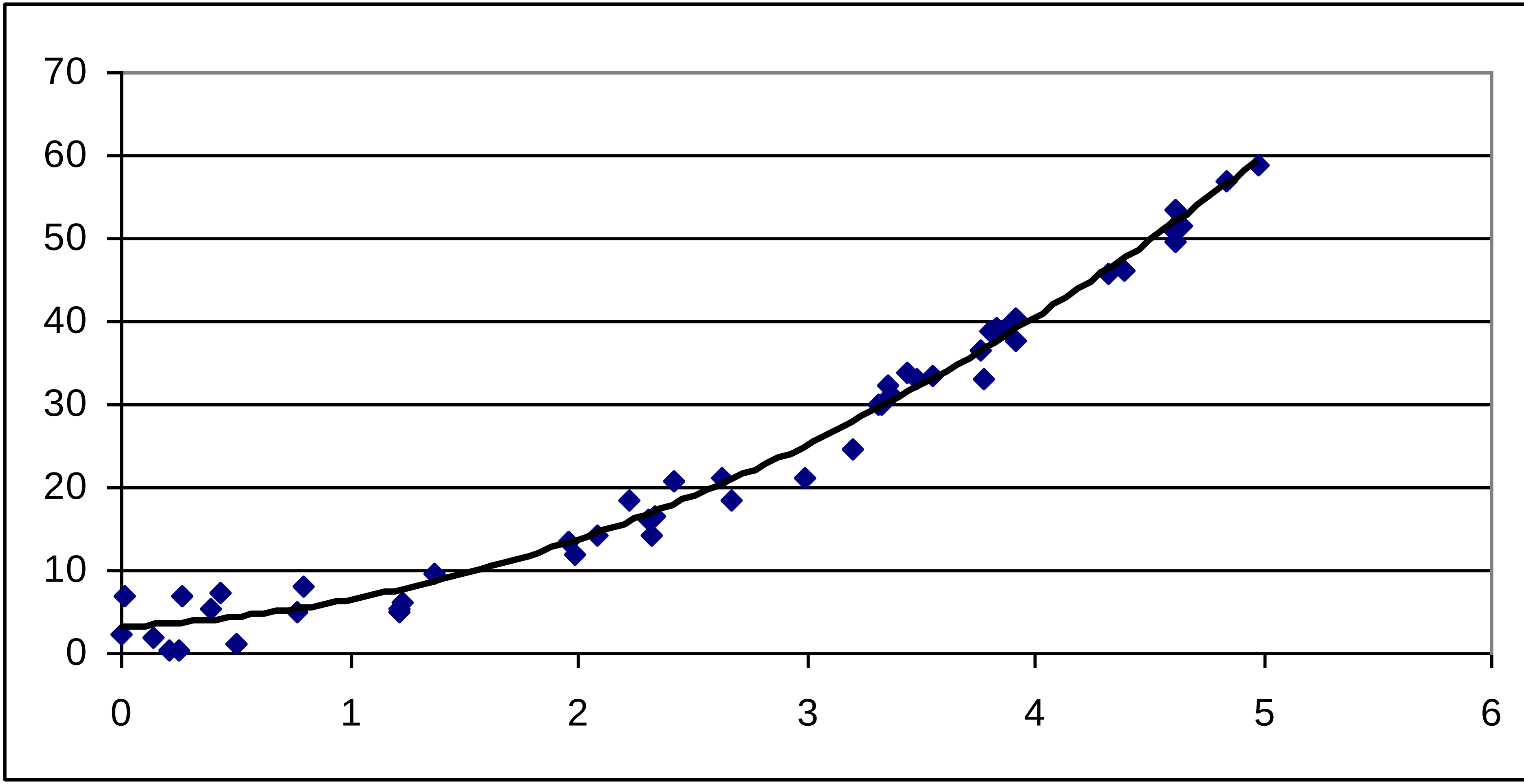
$$E(u \mid x_1, x_2, \ldots, x_k) = 0$$

- For any combination of the independent variables, the expected value of the error term is zero.
- We are equally likely to under-predict as we are to over-predict throughout the multivariate distribution of *x*'s.
- Improperly modeling functional form can cause us to violate this assumption.

Assumption MLR4: Zero Conditional Mean



Assumption MLR4: Zero Conditional Mean



Assumption MLR4: Zero Conditional Mean

- Another common way to violate this assumption is to omit an important variable that is correlated with one of our included variables.
- When x_j is correlated with the error term, it is sometimes called an **endogenous** variable.

Unbiasedness of OLS

- Under assumptions MLR1 through MLR4,

$$E(\hat{\beta}_j) = \beta_j \quad \forall j \in [0, k]$$

- In words: The expected value of each population parameter estimate is equal to the true population parameter.
- It follows that including an irrelevant variable, β_n=0 in a regression model does not cause biased estimates. Like the other variables, the expected value of that parameter estimate will be equal to its population value, 0.

Unbiasedness of OLS

- Note: none of the assumptions 1 through 4 had anything to do with the distributions of y or x.
- A non-normally distributed dependent (y) or independent (x) variable does **not** lead to biased parameter estimates.

Omitted Variable Bias

- Recall that omitting an important variable can cause us to violate assumption MLR4. This means that our estimates may be biased.
- How biased is it?
- Suppose the true model is the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- But we only estimate the following:

$$y = \beta_0 + \beta_1 x_1 + u$$

Omitted Variable Bias

- Why would we do that? Unavailability of the data, ignorance . . .
- Wooldredge (pp. 89-91) shows that the bias in β_1 in the second equation is equal to:

$$E(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$$

- Where $\tilde{\delta}_1$ refers to slope in the regression of x_2 on x_1. This indicates the strength of the relationship between the included and excluded variables.

Omitted Variable Bias

- It follows that there is no omitted variable bias if there is no correlation between the included and excluded variables.
- The sign of the omitted variable bias can be determined from the correlation of x_1 and x_2 and the sign of β_2.
- The magnitude of omitted variable bias depends on how important the omitted variable is (size of β_2), and the size of the correlation between x_1 and x_2.

Omitted Variable Bias, example

- Suppose we wish to know the effect of arrest on high school gpa. Suppose it is a simple world in which the true equation is as follows:

$$gpa = \beta_0 + \beta_1 arrest_1 + \beta_2 sc_2 + u$$

- Where *sc* refers to self-control. Unfortunately, we are using a dataset without a good measure of self-control. So instead, we estimate the following model:

$$gpa_i = 2.9 - .3 arrest_i + u_i$$

Omitted Variable Bias, example

$$gpa_i = 2.9 - .3arrest_i + u_i$$

- This model has known omitted variable bias because self-control is not included. What is the direction of the bias?
- The correlation between arrest and self-control is expected to be negative.
- The expected sign of self-control is postive. Students with poor self-control get lower grades.
- So $\beta_2 \tilde{\delta}_1$ is negative, and likely fairly large. Our estimate of the effect of arrest on gpa is too negative (biased) because self-control affects both arrest and gpa.

Omitted Variable Bias, example 2

- Let’s say that the “true” model for state-level homicide uses poverty and female-headed household rates:

```
. reg homrate poverty fem_hh

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   30.10
       Model |  182.681499     2  91.3407495           Prob > F      =  0.0000
    Residual |    142.6035    47  3.03411701           R-squared     =  0.5616
-------------+------------------------------           Adj R-squared =  0.5429
       Total |  325.284999    49  6.63846936           Root MSE      =  1.7419

------------------------------------------------------------------------------
     homrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |   .1360044   .1051719     1.29   0.202    -.0755742     .347583
      fem_hh |   1.142388   .2190723     5.21   0.000     .7016716    1.583105
       _cons |  -8.487782   1.770978    -4.79   0.000    -12.05053   -4.925034
------------------------------------------------------------------------------
```

Omitted Variable Bias, example 2

- Thus, the “true” effect of poverty on homicide is .136.
- But if we omit female headed households from the model we obtain a much higher estimate of the effect of poverty on homicide (.475).
- This has positive bias because the poverty rate is correlated with the rate of female-headed households, and the relationship between female-headed households and poverty is positive.

Omitted Variable Bias, example 2

- Recall, the bias in our estimate: $E(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$
- B2 is equal to 1.14, and δ1 is equal to .297:

```
. reg fem_hh poverty

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   29.69
       Model |  39.0979508     1  39.0979508           Prob > F      =  0.0000
    Residual |  63.2203985    48  1.31709164           R-squared     =  0.3821
-------------+------------------------------           Adj R-squared =  0.3692
       Total |  102.318349    49  2.08812958           Root MSE      =  1.1476

------------------------------------------------------------------------------
      fem_hh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     poverty |   .2967648   .0544682     5.45   0.000     .1872491    .4062806
       _cons |   6.578087    .678227     9.70   0.000      5.21442    7.941754
------------------------------------------------------------------------------
```

- So the overall bias is .297*1.14=.339
- And the difference between the two estimates is .475-.136 = .339

Assumption MLR5: Homoscedasticity

$$\mathrm{var}(u \mid x_1, x_2, ..., x_j) = \sigma^2$$

- In the multivariate case, this means that the variance of the error term does not increase or decrease with any of the explanatory variables x_1 through x_j.

Variance of OLS estimates

$$\operatorname{var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

- σ^2 is a population parameter referring to error variance. It's an unknown, and something we have to estimate.
- SST_j is the total sample variation in x_j. All else, equal, we would like to have more variation in x, since it means more precise estimates of the slopes. We can get more total sample variation by increasing variation in x or increasing sample size.

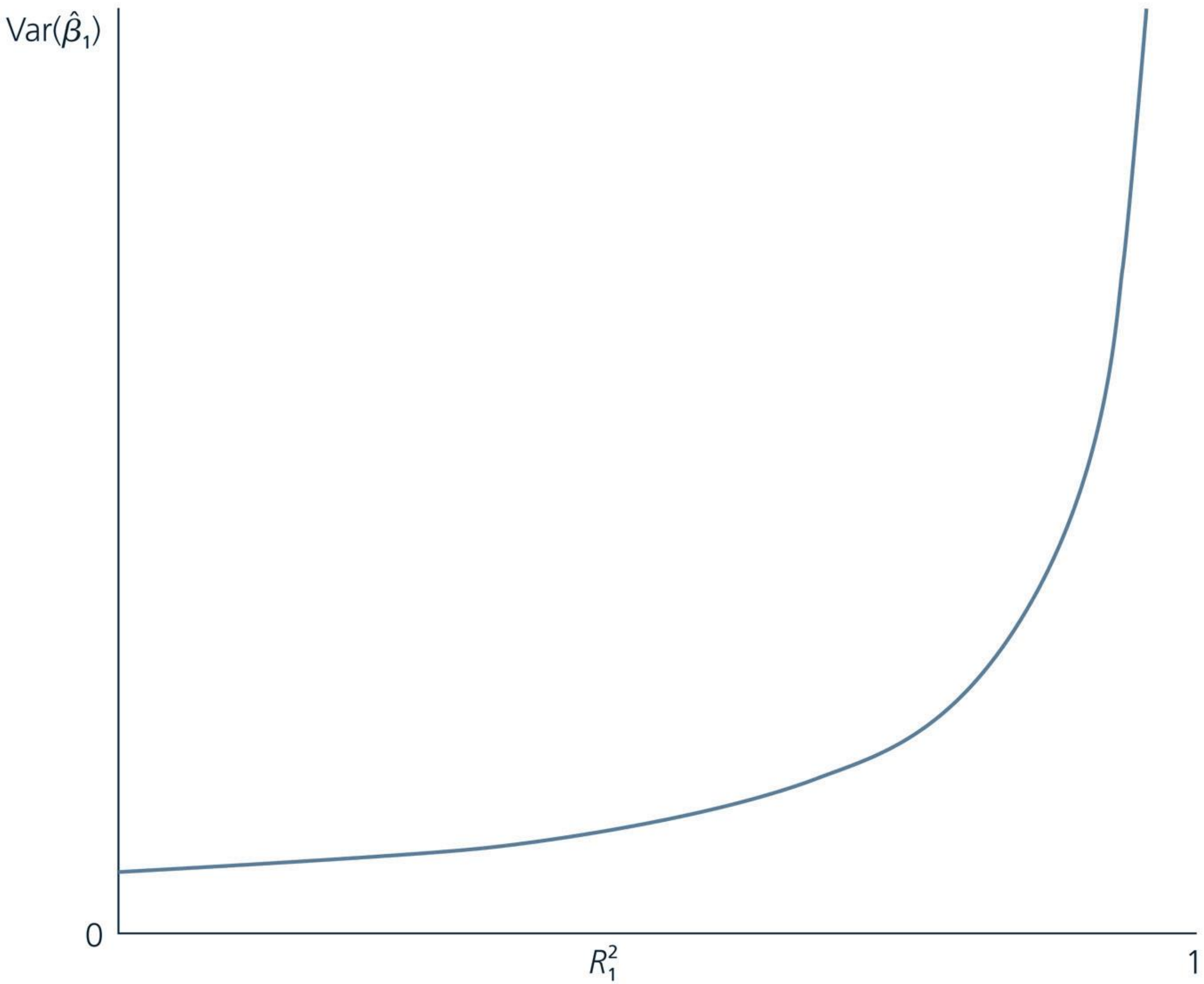
Variance of OLS estimates

$$\operatorname{var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

- R^2_j is the r-squared from the regression of x_j on all other x's.
- This is where multicollinearity comes into play. If there is a lot of multicollinearity, this auxiliary r-squared will be quite large, and this will inflate the variance of the slope estimate.

FIGURE 3.1

$\mathrm{Var}(\hat{\beta}_1)$ as a function of R_1^2.



Variance of OLS estimates

$$\mathrm{var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

- $1/(1-R^2_j)$ is termed the variance inflation factor (VIF). It reflects the degree to which the variance of the slope estimate is inflated due to multicollinearity, compared to zero multicollinearity ($R^2_j = 0$).
- Some researchers have attempted to set up cutoff points above which multicollinearity is a problem. But these should be used with caution.

Variance of OLS estimates

- A high VIF may not be a problem since total variance depends on two other factors, and even very high variance is not a problem if β is relatively much larger.
- You can obtain VIFs using, not surprisingly, the “vif” command after a regression model in Stata.

When OLS is BLUE

- Gauss-Markov Theorem: under assumptions MLR1 through MLR5, OLS estimates are the best (i.e. most efficient) linear unbiased estimates of the population model.

Next time:

Homework: Problems 3.8, 3.10i and ii, C3.8

Read: Wooldridge Chapter 4