CLRM Problems

ME104: Linear Regression Analysis

Kenneth Benoit

August 16, 2012

Classic illustration: the Anscombe dataset

```
. insheet using http://www.kenbenoit.net/courses/quant2/anscombe.csv
(8 vars, 11 obs)

. list, clean

       x1   x2   x3   x4    y1    y2    y3    y4
  1.   10   10   10    8     8   9.1   7.5   6.6
  2.    8    8    8    8   6.9   8.1   6.8   5.8
  3.   13   13   13    8   7.6   8.7    13   7.7
  4.    9    9    9    8   8.8   8.8   7.1   8.8
  5.   11   11   11    8   8.3   9.3   7.8   8.5
  6.   14   14   14    8    10   8.1   8.8     7
  7.    6    6    6    8   7.2   6.1   6.1   5.3
  8.    4    4    4   19   4.3   3.1   5.4    13
  9.   12   12   12    8    11   9.1   8.1   5.6
 10.    7    7    7    8   4.8   7.3   6.4   7.9
 11.    5    5    5    8   5.7   4.7   5.7   6.9
```

Classic illustration: the Anscombe dataset

```
. format x1-y4 %4.2g

. summarize, format

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          x1 |        11           9         3.3          4         14
          x2 |        11           9         3.3          4         14
          x3 |        11           9         3.3          4         14
          x4 |        11           9         3.3          8         19
          y1 |        11         7.5           2        4.3         11
-------------+--------------------------------------------------------
          y2 |        11         7.5           2        3.1        9.3
          y3 |        11         7.5           2        5.4         13
          y4 |        11         7.5           2        5.3         13
```

Classic illustration: the Anscombe dataset

```
. regress y1 x1, cformat(%4.2g)

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.99
       Model |  27.5100011     1  27.5100011           Prob > F      =  0.0022
    Residual |  13.7626904     9  1.52918783           R-squared     =  0.6665
-------------+------------------------------           Adj R-squared =  0.6295
       Total |  41.2726916    10  4.12726916           Root MSE      =  1.2366

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |         .5        .12     4.24   0.002          .23         .77
       _cons |          3        1.1     2.67   0.026          .46         5.5
------------------------------------------------------------------------------
```

Classic illustration: the Anscombe dataset

```
. regress y2 x2, cformat(%4.2g)

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.97
       Model |  27.5000024     1  27.5000024           Prob > F      =  0.0022
    Residual |   13.776294     9  1.53069933           R-squared     =  0.6662
-------------+------------------------------           Adj R-squared =  0.6292
       Total |  41.2762964    10  4.12762964           Root MSE      =  1.2372

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |         .5         .12     4.24   0.002          .23         .77
       _cons |          3         1.1     2.67   0.026          .46         5.5
------------------------------------------------------------------------------
```

Classic illustration: the Anscombe dataset

```
. regress y3 x3, cformat(%4.2g)

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.97
       Model |  27.4700075     1  27.4700075           Prob > F      =  0.0022
    Residual |  13.7561905     9  1.52846561           R-squared     =  0.6663
-------------+------------------------------           Adj R-squared =  0.6292
       Total |  41.2261979    10  4.12261979           Root MSE      =  1.2363

------------------------------------------------------------------------------
          y3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x3 |         .5        .12     4.24   0.002          .23         .77
       _cons |          3        1.1     2.67   0.026          .46         5.5
------------------------------------------------------------------------------
```

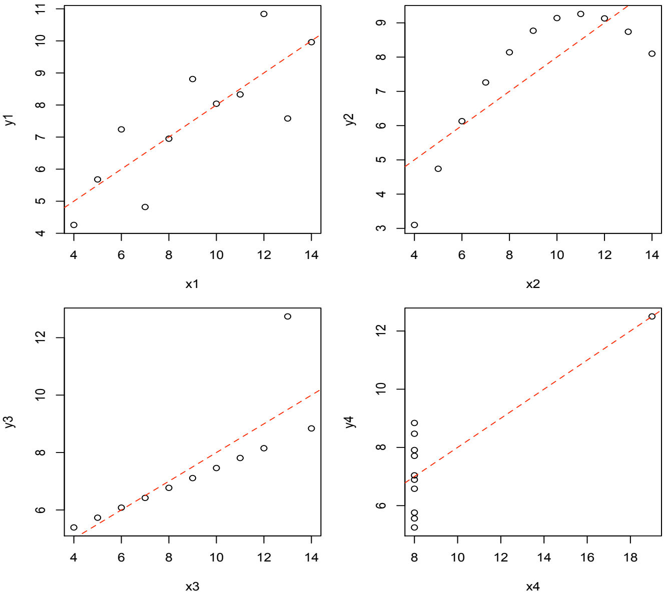
Classic illustration: the Anscombe dataset

```
. regress y4 x4, cformat(%4.2g)

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   18.00
       Model |  27.4900007     1  27.4900007           Prob > F      =  0.0022
    Residual |  13.7424908     9  1.52694342           R-squared     =  0.6667
-------------+------------------------------           Adj R-squared =  0.6297
       Total |  41.2324915    10  4.12324915           Root MSE      =  1.2357

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x4 |         .5        .12     4.24   0.002          .23         .77
       _cons |          3        1.1     2.67   0.026          .46         5.5
------------------------------------------------------------------------------
```

Anscombe dataset plotted



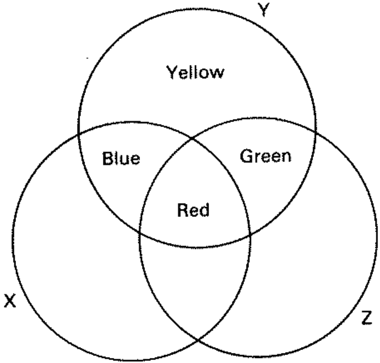
CLRM assumptions revisited

1. Specification:
   - $\text{E}(Y) = X\beta$ (linearity)
   - No extraneous variables in X
   - No omitted independent variables from X
   - Parameters (β) are *constant*
2. $\text{E}(\epsilon) = 0$
3. Error terms:
   - $\text{Var}(\epsilon) = \sigma^2$, or homoskedastic errors
   - $\text{E}(r_{\epsilon_i,\epsilon_j}) = 0$, or no auto-correlation
4. X is non-stochastic
   - implies no *measurement error* in X
   - implies no serial correlation where a lagged value of Y would be used as an independent variable
   - no *simultaneity* or *endogenous* X variables
5. $\text{rank}(X) = k$
6. $\epsilon | X \sim N(0, \sigma^2)$

Omitting a relevant independent variable

- In general, β^{OLS} of included coefficients will be biased, unless the excluded variable is uncorrelated with the included independent variables
- If excluded variable is *orthogonal* to included variables, then β^{OLS} unbiased but α^{OLS} (intercept) wil be biased unless mean of excluded variable is zero
- Variance-covariance matrix of β^{OLS} will be smaller, meaning the MSE of β^{OLS} can go up or down (depending on bias)
- Estimate of var-covariance matrix of β^{OLS} is biased upward, because $\hat{\sigma}^2$ is biased upward, so inferences are inaccurate

Omitting a relevant variable Z: graphical intuition



- Only blue and red areas reflect information used to estimate β in Y on X, but red also reflects variation in Z
- If Z were included, only blue area would be used to estimate β
- Only yellow is used to estimate σ^2, except when Z excluded, and then green area is also used
- If X is orthogonal to Z, then no red area and bias disappears

Including an irrelevant independent variable

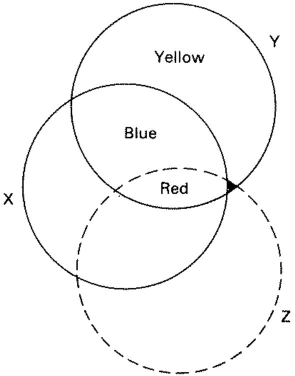
- β^{OLS} and the estimator of its variance-covariance matrix will remain unbiased
- Generally the variance-covariance of β^{OLS} will become larger, and therefore β^{OLS} will be less efficient (increases MSE)
- Change in effect of s_{b_1} of including irrelevant x_2:

$$s_{b_1} = \frac{\hat{\sigma}}{\sqrt{\sum(X_1 - \bar{X}_1)(1 - R^2)}}$$

  so adding another variance will increase R^2 (unless $r_{x_1,x_2} = 0$)

- Keep in mind that "relevant" is a very substantive matter

Adding an irrelevant variable Z: graphical intuition



- Blue area refects variation in Y due entirely to X, so β unbiased
- Since blue area $<$ (blue+red) area, var($\hat{\beta}$) increases
- Yellow area used to estimate σ unbiased so var-cov matrix of $\hat{\beta}$ remains unbiased
- If Z is orthogonal to X then no red area and then no efficiency loss

Non-linearity

- Some non-linear forms simply cannot be used with OLS
- But others can be, if the transformation of one or more variables results in a linear function in the transformed variables
- Two types of transformations, depending on whether the whole equation or only independent variables are transformed
- Transforming only the independent variables example:

$$\begin{aligned} y &= \alpha + \beta_1 x + \beta_2 x^2 + \epsilon \\ y &= \alpha + \beta_1 x + \beta_2 z + \epsilon \end{aligned}$$

  where a new variable $z = x^2$ is created from squaring x
- The equation with z is linear in the parameters but not in the variables

Non-linearity

- Transformating the entire equation means applying a transformation to both sides, not just the independent variables
- Example: the Cobb-Douglas production function:

$$
\begin{aligned}
Y &= AK^{\alpha}L^{\gamma}\epsilon \\
\ln Y &= \ln A + \alpha \ln K + \gamma \ln L + \ln \epsilon \\
Y^* &= A^* + \alpha K^* + \gamma L^* + \epsilon^*
\end{aligned}
$$

is now linear in the transformed variables Y^*, K^* and L^*.

Functional forms for additional non-linear transformations

log-linear as with the Cobb-Douglas production function example

semi-log has two forms:

- $Y = \alpha + \beta \ln X$ (where β is ΔY due to $\%\Delta X$)
- $\ln Y = \alpha + \beta X$ (where β is $\%\Delta Y$ due to ΔX)

inverse or reciprocal: $Y = \alpha + \beta(1/X)$

polynomial $Y = \alpha + \beta X + \gamma X^2$

logit $y = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$ constrains y to lie in $[0, 1]$. Estimation is done by transforming y into log-odds ratio $\ln[y/(1-y)] = \alpha + \beta x$

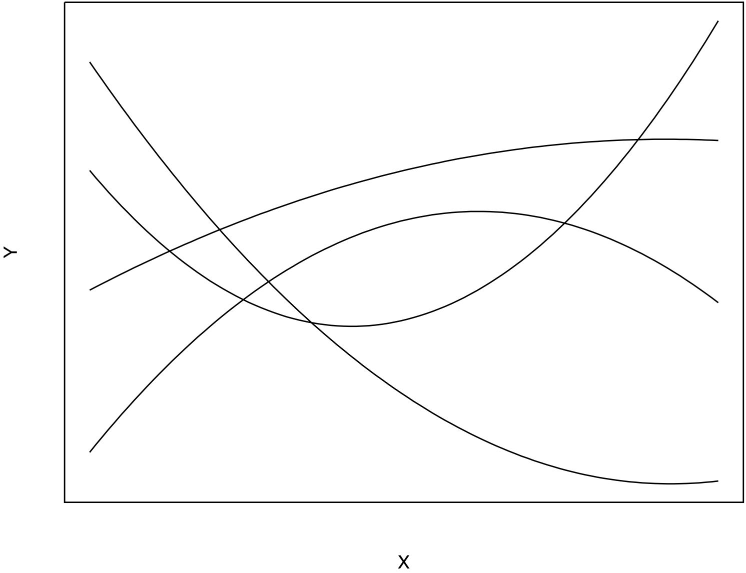
Nonlinear functions of explanatory variables

- A linear regression model can also include explanatory variables which are actually nonlinear transformations of initial explanatory variables
- This means that their association with the response variable does not need to be described by a straight line
- A common example are *polynomial* regression models, in particular the quadratic model

$$E(Y) = \alpha + \beta_1 X + \beta_2 X^2$$

  - which can also include other explanatory variables, here omitted
- This can describe various kinds of nonlinear relationships (see next page)

Nonlinear functions of explanatory variables



Example of a quadratic model

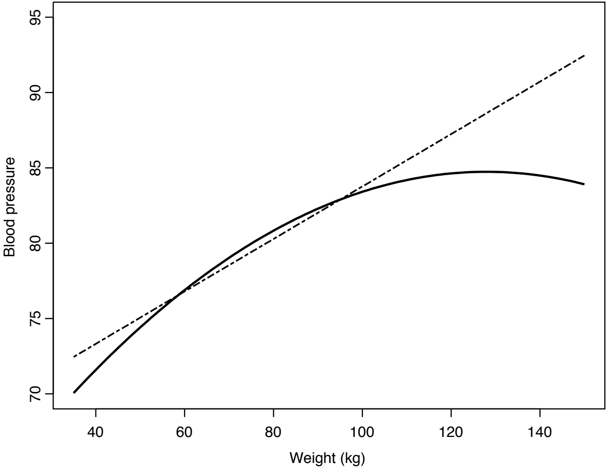
- From HIE data, for blood pressure at exit, given initial blood pressure and
  - respondent's weight: only a linear effect of weight, or
  - both weight and weight^2: a nonlinear (quadratic) effect of weight
- The coefficient of weight^2 is significant at the 5% level ($P = 0.023$), so the quadratic model is preferred
- Nonlinear effects are easiest to interpret using fitted values: see the plot below

Example of a quadratic model

Response variable: diastolic blood pressure at exit				
	Effect of weight			
Variable	Linear		Quadratic	
(Constant)	27.36		18.06	
Initial blood pressure	0.520	(< 0.001)	0.518	(< 0.001)
Weight	0.174	(< 0.001)	0.435	(< 0.001)
Weight^2	—		-0.0017	(0.023)

(P-values in parentheses)

Example of a quadratic model



(Initial blood pressure fixed at 75.)

Logarithms of explanatory variables

- Another common nonlinear transformation of explanatory variables is to use logarithms of them
  - In particular, often used for variables with very skewed distributions
- Leads to linear models of the form

$$E(Y) = \alpha + \beta \log(X)$$

  (usually including other explanatory variables as well)
- The coefficient β of $\log(X)$ is interpreted in terms of proportional changes in X:
  - β is the expected change in Y when X is multiplied by 2.72, i.e. increases by 172%
  - 0.095β is the expected change in Y when X is multiplied by 1.1, i.e. increases by 10%

Example from HIE data

- Response variable: diastolic blood pressure at exit
- Explanatory variables:
  - Initial blood pressure, age, sex, free health care
  - Log of $(1+)$ annual family income
- The estimated coefficient of log-income is -1.298
  - Thus the estimated effect of a 10%-increase in family income is a $0.095 \times 1.298 = 0.123$-point decrease in expected blood pressure, controlling for the other four explanatory variables

Example from HIE data

Variable	Coefficient	P-value
(Constant)	43.99	
Initial blood pressure	0.485	(< 0.001)
Age	0.268	(< 0.001)
Sex: male	4.097	(< 0.001)
Free health care	-1.610	(0.010)
Log of family income	-1.298	(0.007)

Changing parameter values

- No real OLS solutions to this problem in the manner of previous solutions (through transformation)
- For simple "switching regimes" it is possible to divide a dataset into discrete sections, and regress using dummy variables
- A test is available for this, known as the Chow test
- For more complicated and more general models, we must use maximum-likelihood or (even better) Bayesian models
- Example:

$$\begin{aligned} y &= \beta_1 + \beta_2 x + \epsilon \\ \text{where: } \beta_2 &= \alpha_1 + \alpha_2 z + \nu \\ \text{combine to get: } y &= \beta_1 + \alpha_1 x + \alpha_2 (xz) + (\epsilon + x\nu) \end{aligned}$$

Interactions

- There is an interaction between two explanatory variables, if the effect of (either) one of them on the response variable depends on *at which value* the other one is controlled
- Included in the model by using products of the two explanatory variables as additional explanatory variables in the model
- Example: data for the 50 United States, average SAT score of students (Y) given school expenditure per student (X) and % of students taking the SAT in three groups (low, middle and high)
  - The %-variable included as two dummy variables, say D_M for middle and D_L for low

Interactions

- A model without interactions:

$$E(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X$$

- Here the partial effect of expenditure is β_3, same for all values of the %-variable
- Add now the products $(D_L X)$ and $(D_M X)$, to get the model

$$E(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X + \beta_4 (D_L X) + \beta_5 (D_M X)$$

- This model states that there is an interaction between school expenditure and the %-variable
  - Why?

Interactions

- Consider the effect of X at different values of the dummy variables:

$$
\begin{aligned}
&\mathrm{E}(Y) \\
&= \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X + \beta_4 (D_L X) + \beta_5 (D_M X) \\
&\quad = \quad \alpha + \beta_3 X && \text{For high-\% states} \\
&\quad = \quad (\alpha + \beta_2) + (\beta_3 + \beta_5) X && \text{For mid-\% states} \\
&\quad = \quad (\alpha + \beta_1) + (\beta_3 + \beta_4) X && \text{For low-\% states}
\end{aligned}
$$

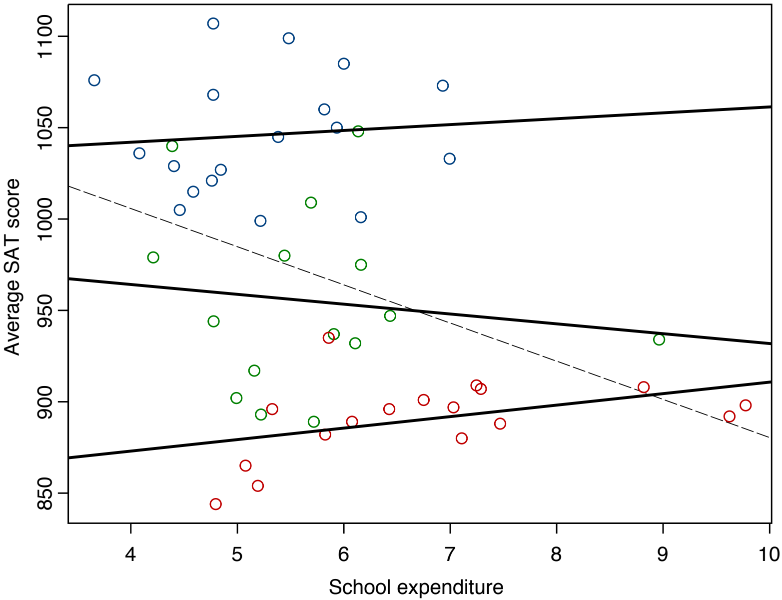
- In other words, the coefficient of X depends on the value at which D_L and D_M are fixed

Interactions

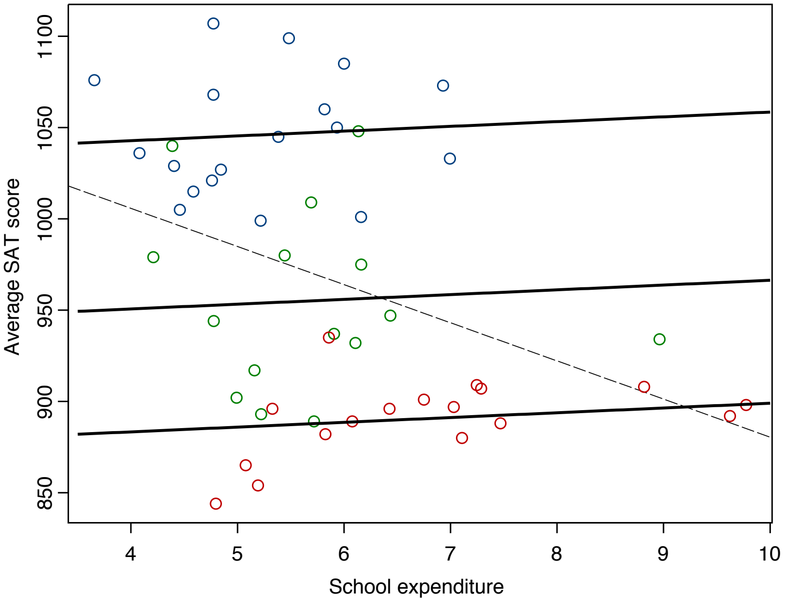
- The estimated coefficients in this example are

$$
\begin{aligned}
\mathrm{E}(Y) &= 847.9 + 181.3D_L + 137.8D_M + 6.3X \\
&\qquad -3.2(D_L X) - 11.7(D_M X) \\
&= 847.9 + 6.3X && \text{for high-\% states} \\
&= 1029.2 + 3.1X && \text{for low-\% states} \\
&= 985.7 - 5.4X && \text{for mid-\% states}
\end{aligned}
$$

Model with interaction



...and without



Average SAT score
850
900
950
1000
1050
1100
4
5
6
7
8
9
10
School expenditure

Testing for interactions

- A standard test of whether the coefficient of the product variable (or variables) is zero is a test of whether the interaction is needed in the model
  - t-test or (if more than one product variable) F-test
- In the example, we use an F-test, comparing

$$\begin{aligned} \text{Full model} \quad & \mathrm{E}(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X \\ & \qquad + \beta_4 (D_L X) + \beta_5 (D_M X) \\ \text{vs. Restricted m.} \quad & \mathrm{E}(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X \end{aligned}$$

  i.e. a test of $H_0: \ \beta_4 = \beta_5 = 0$
- Here $F = 0.61$ and $P = 0.55$, so the interaction is not in fact significant

Interactions between categorical variables

- In the previous example, the interaction was between a continuous variable and a categorical variable
- In other cases too, interactions are included as products of variables
  - For an example of an interaction between two continuous variables, see S. 4.6.2
- An example of interaction between two categorical (here binary) explanatory variables, from HIE data:
  - Response variable: blood pressure at exit
  - Two binary explanatory variables:
    - Being on free health care vs. some other plan
    - Income in the lowest 20% in the data vs. not
  - Other control variables: initial blood pressure, age and sex

Interactions between categorical variables

Variable	Coefficient
Initial blood pressure	0.483
Age	0.260
Sex: Male	3.981
Low income (lowest 20%)	2.662
Free health care	-1.299
Income×Insurance plan	-1.262
(Constant)	31.83

Interactions between categorical variables

- Which coefficients involving income and insurance plan apply to different combinations of these variables:

Free care	Low income: No	Low income: Yes
No	0	2.662
Yes	-1.299	0.101

(not showing the other coefficients)

  where 0.101=2.662-1.299-1.262
- In other words,
  - effect of low income on blood pressure is smaller for respondents on free care than on other plans
  - effect of free care on blood pressure is bigger for low-income respondents than for high-income ones
- (Again, the interaction is not actually significant ($P = 0.42$) here, so this just illustrates the general idea)