Analysis of Economics Data
Chapter 16: Checking the Model and Data

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 16: Checking the Model and Data

- We assume the data are such that
  - There is **variation in the sample regressors** so that the regressors are not perfectly correlated with each other.
- Analysis under the strongest assumptions 1-4 assumes that in the population
  - **1.** The **population model** is $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$.
  - **2.** The **error has mean zero conditional on x**: $\mathrm{E}[u_i | x_{2i}, ..., x_{ki}] = 0$.
  - **3.** The **error has constant variance conditional on x**: $\mathrm{Var}[u_i | x_{2i}, ..., x_{ki}] = \sigma_u^2$.
  - **4.** The **errors for different observations are statistically independent**: u_i is independent of u_j.
- What happens if one or more of these assumptions fail?
- Also consider influential and outlying observations.

Outline

Datasets: EARNINGS_COMPLETE, DEMOCRACY

16.1 Multicollinear Data

- We need sufficient variation in the regressors.
- Extreme case is **perfect collinearity**
  - then not all coefficients can be estimated
  - e.g. dummy variable trap where $d1 + d2 + d3 = 1$ and include all three.
- More generally problem is there is very high correlation between the regressors
  - Then OLS is still unbiased and consistent
  - but individual coefficients may be very imprecisely estimated
- Example is earnings regression with regressors age *(Age)*, years of education *(Education) and years of work experience (Experience)*
  - expect that $Experience \simeq Age - Education - 6$
  - it will be difficult to disentangle the separate roles of age, education and years of work experience.

Multicollinearity: Detection and Solution

- OLS is still unbiased and consistent with multicollinearity
  - and prediction is still okay
  - problem is imprecise estimation of individual coefficients.
- Detection
  - Signs of multicollinearity are high standard errors, low t-statistics and "wrong" signs.
  - A simple diagnostic method is to regress one regressor on the remaining regressors
    - if R^2 is very high then multicollinearity is a problem
    - if $R^2 = 1$ then there is perfect collinearity.
  - Note: can have multicollinearity even if pairwise correlations are small.
- Solution
  - get more data
  - drop one or more variables
  - if subset is collinear just do joint F test on this subset.

16.2 Model Assumptions Revisited

- Recall assumptions 1-4 (for bivariate model for simplicity)
  - **1.** The **population model** is $y_i = \beta_1 + \beta_2 x_i + u_i$ for all i.
  - **2.** The **error for the i^{th} observation has mean zero conditional on x**: $\mathrm{E}[u_i|x_i] = 0$ for all i.
  - **3.** The **error for the i^{th} observation has constant variance conditional on x**: $\mathrm{Var}[u_i|x_i] = \sigma_u^2$ for all i.
  - **4.** The **errors for different observations are statistically independent**: u_i is independent of u_j for all $i \neq j$.
- Failure of assumptions 1 and/or 2
  - OLS biased and inconsistent
- Failure of assumptions 3 and/or 4 (but 1 and 2 okay)
  - OLS unbiased and consistent
  - But different standard errors than the default
  - So default gives invalid confidence intervals and t-statistics.

Why do the Assumptions Matter?

- Appendix C.1 provides full details. Here just a summary.
- Consider regression of y on just x (no intercept)
  - so $b = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i{}^2$
  - or $b = \sum_{i=1}^{n} w_i y_i$ where $w_i = x_i / \sum_{i=1}^{n} x_i{}^2$
- First we need to specify a model for y_i
  - If $y_i = \beta x_i + u_i$ (assumption 1)
  - then some algebra shows $b = \beta + \sum_{i=1}^{n} w_i u_i$.
- Next for b to be unbiased for β we need $\mathrm{E}[\sum_{i=1}^{n} w_i u_i] = 0$
  - this is the case if $\mathrm{E}[u_i|x_i] = 0$.
- Next given the above
  - $\mathrm{Var}[b] = \mathrm{E}[(b-\beta)^2] = \mathrm{E}\left[\left(\sum_{i=1}^{n} w_i u_i\right)^2\right]$
  - Under assumptions 3 and 4 this gives $\mathrm{Var}[b] = \sigma_u^2 / \left(\sum_{i=1}^{n} x_i{}^2\right)$.

16.3 Incorrect Population Model

- The population model is no longer $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$.
- Wrong functional form - e.g. linear-log and not linear
  - OLS is biased and inconsistent
- Omitted regressors - model should have included additional regressors
  - OLS is biased and inconsistent (if the additional regressors are correlated with the included regressors).
- Omitted variables bias
  - True model: $y = \alpha_1 + \alpha_2 x + \alpha_3 z + v$
  - Estimated model: $y = \beta_1 + \beta_2 x + u$ (so z is omitted)
  - Then $\mathrm{E}[b_2] = \alpha_2 + \alpha_3 \times \gamma$ where $\gamma = \Delta z / \Delta x$ is coefficient form OLS of z on x.
- Irrelevant regressors - some of the regressors should not have been included
  - OLS is unbiased and consistent but not as precisely estimated
  - so better to have too many regressors than too few.

16.4 Regressors Correlated with Errors

- Regressors are correlated with the errors
- For simplicity consider $y = \beta_1 + \beta_2 x + u$
  - Problem is $\mathrm{E}[u|x] \neq 0$ so OLS is biased and inconsistent
  - e.g. (1) u includes omitted variables that are correlated with x
    - ★ y is earnings, x is schooling and u includes unobserved ability
  - e.g. (2) feedback from y to x
    - ★ y is inflation and x is money supply growth.
- General term is that regressor x correlated with u is endogenous.
- One solution is instrumental variables estimation assuming an instrument exists
  - instrument is uncorrelated with u (so does not determine y)
  - but correlated with x.

16.5 Heteroskedastic Errors

- Now $\text{Var}[u_i | x_{2i}, ..., x_{ki}]$ varies with i.
- Heteroskedastic errors
  - common for cross-section data independent across observations
- Usual response is to do OLS but base inference on heteroskedastic-robust standard errors.
- In some cases transform y so that error is less heteroskedastic
  - e.g. log-earnings regressions
- In some cases provide a model for the heteroskedasticity and estimate by feasible generalized least squares.

16.6 Correlated Errors

- Now u_i is correlated with u_j
  - two leading examples - autocorrelated errors and clustered errors
- Clustered errors arise with (short) panel data and some cross-section data
  - errors in same cluster (e.g. village) are correlated with each other
  - do OLS but get cluster-robust standard errors
  - or assume e.g. random effects model and estimate by feasible generalized least squares
  - chapter 17.
- Autocorrelated errors arise with time series data
  - e.g. $u_t = \rho u_{t-1} + \varepsilon_t$ (autoregressive error of order 1 or AR(1))
  - do OLS but get heteroskedastic and autocorrelation consistent (HAC) standard errors
  - or assume e.g. AR(1) error and estimate by feasible generalized least squares
  - chapter 17.

16.7 Example: Democracy and Growth

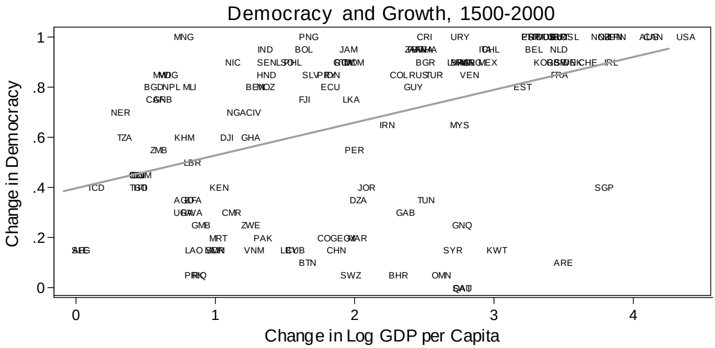
- Dataset DEMOCRACY has data for 131 countries from Daron Acemoglu, Simon Johnson, James A. Robinson, and Pierre Yared (2008), "Income and Democracy," American Economic Review, Vol.98, pp. 808-42.

Variable	**Definition**	**Mean**	**Standard Deviation**	**I**
Democracy	500 year democracy change (1500-2000)	0.647	0.3310	
Growth	500 year income per capita change (1500-2000)	1.916	1.108	-0
Constraint	Constraint on the executive at independence	0.372	0.3622	
Indcent	Year of independence / 100	19.044	0.677	1
Catholic	Catholics proportion of population in 1980	0.305	0.355	
Muslim	Muslim proportion of population in 1980	0.250	0.371	
Protestant	Protestant proportion of population in 1980	0.127	0.213	
Other	Other religion proportion of population in 1980	0.320	0.320	0
n	131			

Bivariate Regression

- OLS (with heteroskedastic-robust standard errors)

$$\widehat{Democracy} = \underset{(0.046)}{0.397} + \underset{(0.019)}{0.131}\,Growth, \quad R^2 = .192, \quad n = 131.$$



Multiple Regression

- OLS (with heteroskedastic-robust standard errors)

$$\begin{aligned}\widehat{Democracy} &= \underset{(0.870)}{3.031} + \underset{(0.025)}{0.047}\,Growth + \underset{(0.072)}{0.164}\,Constraint - \underset{(0.050)}{0.133}\,Indce\\ &\quad + \underset{(0.089)}{0.117}\,Catholic - \underset{(0.101)}{0.233}\,Muslim + \underset{(0.180)}{0.180}\,Protestant,\\ &\quad R^2 = .192, \quad n = 131.\end{aligned}$$

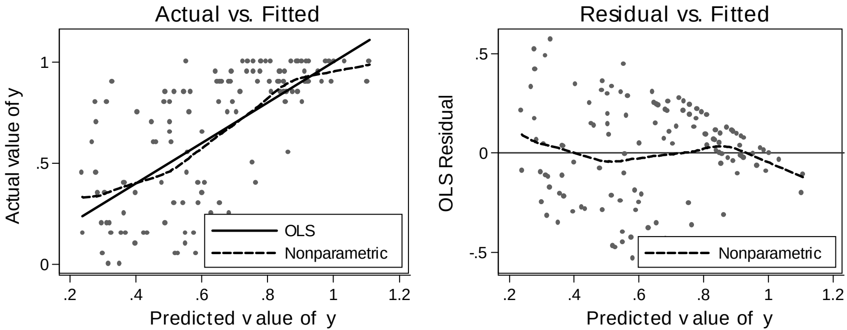
- Coefficient of *Growth* fell from 0.131 to 0.047
  - point of article is that institutions such as religion matter
  - here higher for Catholic and Protestant than for Muslim and Other religions.

16.8 Diagnostics: Outliers and Influential Observations

- An **outlier** or **outlying observation** is one whose value is unusual given the rest of the data.
- Need to screen for these as may be due to erroneous data.
- Also outlier may have large effect on results of OLS estimation
  - bivariate if (x_i, y_i) a long way from $(\bar{x}, \bar{y})$
  - since $b_2 = \left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right] / \left[\sum_{i=1}^n (x_i - \bar{x})^2\right]$
- For multiple regression an **influential observation** is one that has a relatively large effect on the results of regression analysis, notably on $\widehat{y}$ or on estimated OLS coefficients.
- Not all outliers are influential observations.
  - An outlier with regressor value a long way from the sample mean $\bar{x}$ is said to have high **leverage**.

Scatter Plots against the Fitted Values

- First panel: plot y against $\widehat{y}$ shows nothing systematically wrong
- Second panel: plot $e = y - \widehat{y}$ against $\widehat{y}$ (rotates the first figure).

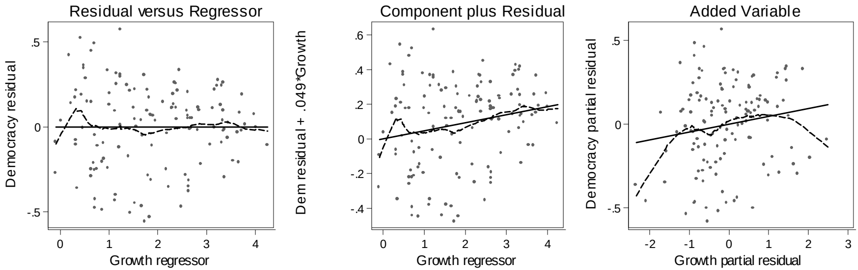


Scatter Plots for a Single Regressor

Panel 1: **residual versus regressor plot:** plot e_i against x_{ji}

Panel 2: **component plus residual plot** or **partial residual plot** is a plot of $p_{ji} = b_j x_{ji} + e_i$ against x_{ji}

Panel 3: **added variable plot** or **partial regression leverage plot** is plot of y against x_j after purging both y and x_j of the effect of the other regressors.



Detecting Influential Observations

- DFITS measures the influence of a particular observation on the fitted values.
- DFITS_i equals the scaled difference between predictions of y_i with and without the i^{th} observation included in the OLS regression (so **DFITS** means **difference in fits**).
  - Large absolute values of DFITS indicate an influential observation
  - conservative rule of thumb is suspicious observations have $|\text{DFITS}| > 2\sqrt{k/n}$, where k is # regressors and n is sample size
- DBETA measures the influence of a particular observation on the coefficients.
- For j^{th} regressor and i^{th} observation DFBETA_{ji} equals the scaled difference between b_j with and without the i^{th} observation included in the OLS regression (so **DFBETA** means **difference in beta**).
  - Conservative rule of thumb is that observations with $|\text{DFBETA}| > 2/\sqrt{n}$ may be worthy of further investigation.

Residual Distribution

- Residuals that are unusually large in absolute values may indicate outliers.
- Asymmetric residuals may indicate that a nonlinear model needs to be estimated.
- But note that residuals are not the same as model errors

$$\begin{aligned} e_i &= y_i - b_1 - b_2 x_i \\ &= y_i - \beta_1 - \beta_2 x_i - b_1 + \beta_1 - b_2 x_i + \beta_2 x_i \\ &= u_i - (b_1 - \beta_1) - (b_2 - \beta_2) x_i, \end{aligned}$$

  using $y_i = \beta_1 + \beta_2 x_i + u_i$.
- So e_i depends on x_i (and on other $x's$ through estimates b_1 and b_2) even if u_i does not.
  - This dependence disappears as $n \rightarrow \infty$ since $(b_1 - \beta_1) \rightarrow 0$ and $(b_2 - \beta_2) \rightarrow 0$.
  - But in finite samples residuals are heteroskedastic and correlated even if model errors are not.

Key Stata Commands

```
regress democracy growth constraint indcent ///
    catholic muslim protestant
* Residual versus a regressor plot
rvpplot growth, yline(0)
* Component plus residual plot
cprplot growth, lowess
* Added variable plot
avplot growth
* Influential observations
predict dfits, dfits
predict dfbgrowth, dfbeta(growth)
```

Some in-class Exercises

1. We estimate by OLS the model $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ and obtain default standard errors. What problems arise when, in turn, each of the following occurs.
    1. x_3 should not appear in the model.
    2. x_3 is an indicator variable that takes only values 0 or 1.
    3. $x_3 = 2x_2$.
    4. x_4 should also have appeared in the model.
    5. u_i has mean zero but it is not independent of the other u_j.
    6. u_i has have mean zero and is independent of the other u_j, but it is heteroskedastic.