Diagnostics of Linear Regression

Junhui Qian

October 27, 2014

The Objectives

- After estimating a model, we should always perform diagnostics on the model. In particular, we should check whether the assumptions we made are valid.
- For OLS estimation, we should usually check:
  - Is the relationship between x and y linear?
  - Are the residuals serially uncorrelated?
  - Are the residuals uncorrelated with explanatory variables? (endogeneity)
  - Does homoscedasticity hold?

Residuals

- Residuals are unobservable. But they can be estimated:

$$\hat{u}_i = y_i - x_i'\hat{\beta}.$$

- Using matrix language,

$$\hat{u} = (I - P_X)Y.$$

- If $\hat{\beta}$ is close to β, then $\hat{u}_i$ is close to u_i.
- Let $\hat{y}_i = x_i'\hat{\beta}$, we call $\hat{y}_i$ the "the fitted value".
- Then the explained variable can be decomposed into

$$y_i = \hat{y}_i + \hat{u}_i.$$

Variations

- SST (total sum of squares)

$$\mathrm{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 = Y'(I - P_\iota)Y.$$

- SSE (explained sum of squares)

$$\mathrm{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = Y'(P_X - P_\iota)Y.$$

- SSR (sum of squared residuals)

$$\mathrm{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2 = Y'(I - P_X)Y.$$

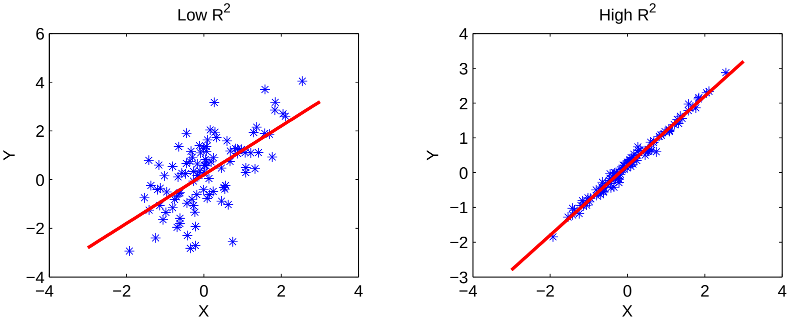
- We have $\quad \mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}.$

Goodness of Fit

- R^2 of the regression:

$$R^2 \equiv \mathrm{SSE}/\mathrm{SST} = 1 - \mathrm{SSR}/\mathrm{SST}.$$

- R^2 is the fraction of the sample variation in y that is explained by x. And we have $0 \leq R^2 \leq 1$.
- R^2 does NOT validate a model. A high R^2 only says that y is predictable with information in x. In social science, this is not the case in general.
- If additional regressors are added to a model, R^2 will increase.
- The adjusted R^2, denoted as $\bar{R}^2$, is designed to penalize the number of regressors,
  $\bar{R}^2 = 1 - [\mathrm{SSR}/(n-1-k)]/[\mathrm{SST}/(n-1)$.


Low R^2
6
4
2
0
−2
−4
−4
−2
0
2
4
Y
X
High R^2
4
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
Y
X

Residual Plots

We can plot

- Residuals
- Residuals versus Fitted Value
- Residuals versus Explanatory Variables

Any pattern in residual plots suggests nonlinearity or endogeneity.

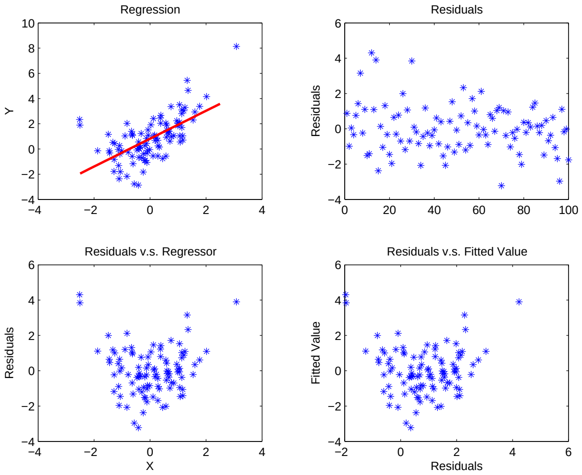


Figure : Residual Plots. DGP: $y = 0.2 + x + 0.5x^2 + u$

Partial Residual Plots

- To see whether there exists nonlinearity in a regressor, say the j-th explanatory variable x_j, We can plot

$$\hat{u} + \hat{\beta}_j x_j \quad \text{versus} \quad x_j,$$

  where $\hat{u}$ is residual from the full model.
- Partial residual plots may help us find the true (nonlinear) functional form of x_j.

Partial Residual Plots: Example

Suppose the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + g(z) + u,$$

where $g(z)$ is a nonlinear function. We mistakenly estimate:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{u}.$$

If we plot $\hat{\beta}_2 z + \hat{u}$ versus z, we may probably be able to detect nonlinearity in $g(z)$.

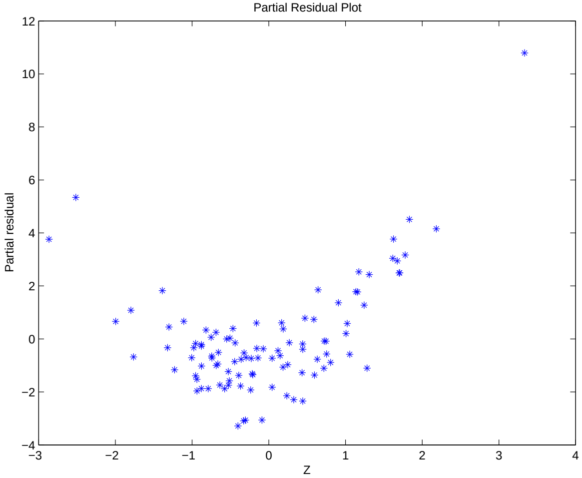


Figure : Residual Plots. DGP: $y = 0.2 + x + 0.5z + z^2 + u$

The iid Assumption

- The CLR assumption dictates that residuals should be iid.
- It is generally difficult to determine whether a given number of observations are from the same distribution.
- If there is a natural order of the observations (e.g., time), then we may check whether the residuals are correlated.
- If there is correlation, then the iid assumption is violated.

Residuals with Time

- When we deal with time series regression, for example,

$$\pi_t = \beta_0 + \beta_0 m_t + u_t,$$

  where π_t is the inflation rate and m_t is the growth rate of money supply, both indexed by time t.
- Now the "natural order" is time, and a time series plot of the estimated residual contains information.

Residual Plots

We can plot:

- Residuals over time
- Residuals v.s. previous residual
- Correlogram

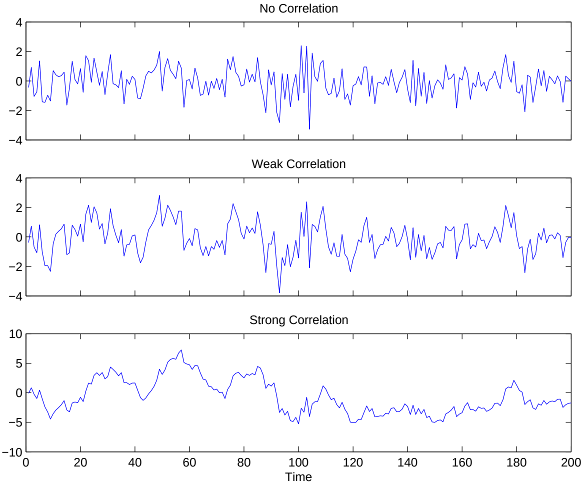


Figure : Residuals over time: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from top to bottom.

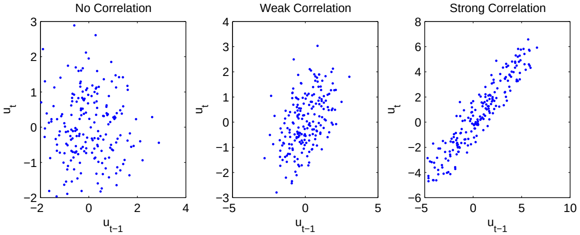


Figure : Residuals v.s. previous residual: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from left to right.

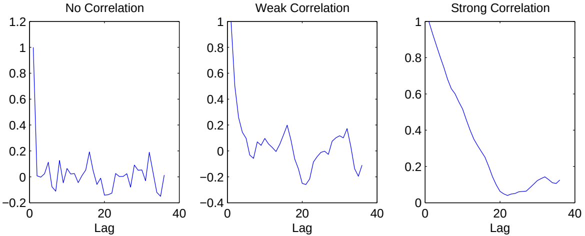


Figure : Correlograms: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from left to right.

Durbin-Watson Test

- Durbin-Watson is the formal test for independence, or more precisely, non-correlation.
- It assumes a AR(1) model for u_t, $u_t = \alpha u_{t-1} + \varepsilon_t$.
- The null hypothesis is: $H_0 : \rho = \alpha = 0$.
- The test statistic is

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_{t-1}^2}.$$

Durbin-Watson Test

- $DW \in [0, 4]$.
- $DW = 2$ indicates no autocorrelation.
- If DW is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if DW is less than 1.0, there may be cause for alarm.
- Small values of DW indicate successive error terms are, on average, close in value to one another, or positively correlated.
- Large values of DW indicate successive error terms are, on average, much different in value to one another, or negatively correlated.

Fixing Correlation

- It's most likely that the model is misspecified.
- The usual practices are:
  - Add more explanatory variables
  - Add more lags of the existing explanatory variables

- If $\mathrm{var}(u_i|x) = \sigma^2$, we call the model "homoscedastic". If not, we call it "heteroscedastic".
- If homoscedasticity does not hold, but CLR Assumptions 1-4 still hold, the OLS estimator is still unbiased and consistent. However, OLS is no longer BLUE.
- We can detect heteroscedasticity by looking at the residuals v.s. regressors.
- For simple regressions, we can look at regression lines.
- And we can formally test for homoscedasticity.
  - White test
  - Breusch-Pagan test

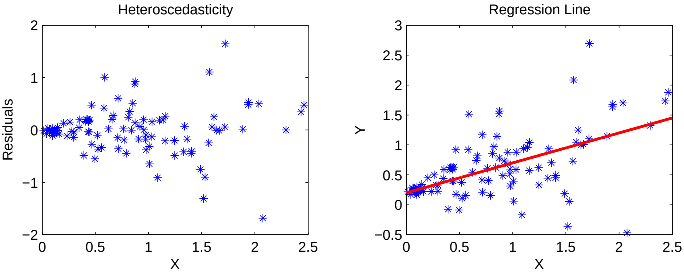


Figure : Heteroscedasticity. DGP: $y_i = \beta_0 + 0.5x_i + x_i\varepsilon_i$.

Fixing Heteroscedasticity

- Use a different specification for the model (different variables, or perhaps non-linear transformations of the variables).
- Use GLS (Generalized Least Square).