Bivariate Regression Discussion

Amir Samimi
Civil Engineering Department
Sharif University of Technology

Primary Source: Basic Econometrics (Gujarati)

Topics

- First we consider the case of regression through the origin:
  - A situation where the intercept term is absent from the model.
- Then we consider the question of the units of measurement.
  - Whether a change in the units of measurement affects the regression results.
- Finally, we consider the question of the functional form of the linear regression model.
  - So far we have considered models that are linear in the parameters as well as in the variables.

Regression Through the Origin

- There are occasions when the PRF assumes the following form: $Y_i = \beta_2 X_i + u_i$
- Sometimes the underlying theory dictates that the intercept term be absent from the model.
  - Cost analysis theory:
    - The variable cost of production is proportional to output.
  - Some versions of monetarist theory:
    - Rate of inflation is proportional to the rate of change of the money supply.
- How do we estimate such models like, and what special problems do they pose?

Estimation

- Write the SRF as: $Y_i = \hat{\beta}_2 X_i + \hat{u}_i$
- Applying the OLS method:
  \[
  \hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2}, \quad \text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_i^2}, \quad \hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 1}
  \]
- Comparison with formulas obtained with intercept:
  \[
  \hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum \hat{u}_i^2}, \quad \text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum \hat{u}_i^2}, \quad \hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 2}
  \]
Intercept-present Versus Intercept-less

- **Intercept-present:**
  - adjusted (from mean) sums of squares and cross products.
  - df for estimating $\sigma^2$ is $(n - 1)$.
  - $\sum \hat{u}_i$ is always zero.
  - Coefficient of determination is always between 0 and 1.

- **Intercept-less:**
  - raw sums of squares and cross products.
  - df for estimating $\sigma^2$ is $(n - 2)$.
  - $\sum \hat{u}_i$ need not be zero.
  - Coefficient of determination is not always between 0 and 1.

R$^2$ for Interceptless Model

- Compute the raw $R^2$ for such models: $\text{raw } r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$

- Raw $r^2$ is not directly comparable to the conventional $r^2$ value.
- Some authors do not report $r^2$ value for interceptless models.
- Unless there is a very strong priori expectation, one would be well advised to stick to the conventional models.
  - If intercept turns out to be statistically insignificant, we have a regression through the origin.
  - If there is an intercept but we insist on fitting a regression through the origin, we would be committing a specification error, thus violating Assumption 9.

Units of Measurement

- Do the units in which the variables are measured make any difference in the regression results?
- If so, what is the sensible course to follow in choosing units of measurement for regression analysis?

\[
Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \\
Y'_i = w_1 Y_i \\
X'_i = w_2 X_i \\
Y''_i = \hat{\beta}_1 + \hat{\beta}_2 X'_i + \hat{u}'_i \\
Y''''_i = w_1 Y'_i, X''''_i = w_2 X'_i, \text{ and } \hat{u}''''_i = w_1 \hat{u}'_i
\]
Discussion

- One should choose the units of measurement sensibly.
- If the scaling factors are identical ($w_1 = w_2$):
  - the slope coefficient and its standard error remain unaffected.
  - The intercept and its standard error are both multiplied by $w_r$.
- If the $X$ scale is not changed:
  - the slope as well as the intercept coefficients and their respective standard errors are all multiplied by the same factor.
- If the $Y$ scale remains unchanged:
  - ?
- Goodness-of-fit is not affected at all.

Standardized Variables

- Subtract the mean value of the variable from its individual values and divide the difference by the standard deviation of that variable:
  \[ Y_i^* = \frac{Y_i - \bar{Y}}{S_Y} \]
  \[ X_i^* = \frac{X_i - \bar{X}}{S_X} \]
- Instead of running $Y_i = \beta_1 + \beta_2 X_i + u_i$
- We could run $Y_i^* = \beta_1^* + \beta_2^* X_i^* + u_i^*$

Interpretation

- How do we interpret the beta coefficients?
- Note that new variables are measured in standard deviation units.
- If the regressor increases by one standard deviation, on average, the regressand increases by $\beta_2^*$ standard deviation units.

Advantages?

- More apparent if there is more than one regressor.
- We put all regressors on equal basis and compare them directly.
- If a coefficient is larger than another, then …
  - the it contributes more to the explanation of the regressand.

Note 1: For the standardized regression we have not given the $r^2$ value because this is a regression through the origin.

Note 2: There is an interesting relationship between the $\beta$ of the conventional model and the standardized one:

$$ \hat{\beta}_2 = \hat{\beta}_2 \left( \frac{S_Y}{S_X} \right) $$
- $S_X$ = The sample standard deviation of the $X$ regressor,
- $S_Y$ = The sample standard deviation of the regressand.
Functional Forms of Regression Models

- We consider some commonly used regression models that may be nonlinear in the variables but are linear in the parameters:
  - The log-linear models
  - Semilog models
  - Reciprocal models
  - The logarithmic reciprocal models

- What are the special features of each model?
- When are they appropriate?
- How are they estimated?

Log-linear Models

- Consider the following model, exponential regression model:
  \[ Y_i = \beta_1 X_i^\beta e^{\varepsilon_i} \]
  \[ \ln Y_i = \ln (\beta_1 + \beta_2 \ln X_i) + \varepsilon_i \]
  \[ \ln Y_i = \alpha + \beta_2 \ln X_i + \varepsilon_i \]

- Linear in the parameters, linear in the logarithms of the variables, and can be estimated by OLS regression.
- Because of this linearity, such models are called log-log, double-log, or loglinear models.

Semilog Models

- The Log-Lin Model:
  - To measure the growth rate.
  \[ Y_i = Y_0 (1 + r)^t \]
  \[ \ln Y_i = \ln Y_0 + \ln(1 + r) \]
  \[ \ln Y_i = \beta_1 + \beta_2 t \]
  - Regressand is the logarithm of \( Y \) and the regressor is “time”.
  \[ \beta_2 = \frac{\text{relative change in regressand}}{\text{absolute change in regressor}} \]

- A model in which \( Y \) is logarithmic is called a log-lin model.
- A model in which \( Y \) is linear but \( X(s) \) are logarithmic is a lin-log model.

Discussion

- One attractive feature of the log-log model, which has made it popular in applied work:
  - \( \beta_2 \) measures the elasticity of \( Y \) with respect to \( X \) (why?).
  - The model assumes that the elasticity coefficient, remains constant throughout, hence called the constant elasticity model.
  - Although \( \alpha \) and \( \beta_2 \) have unbiased estimators, \( \beta_1 \) has not!
    - In most practical problems, \( \beta_1 \) is of secondary importance.
  - How to decide?
    - The simplest way is to plot the scattergram of \( \ln Y_i \) against \( \ln X_i \) and see if the scatter points lie approximately on a straight line.
Semilog Models

- The Lin–Log Model:
  - We want to find absolute change in $Y$ for a percent change in $X$.
    \[ Y_t = \beta_1 + \beta_2 \ln X_t + u_t \]
    \[ \beta_2 = \frac{\text{change in } Y}{\text{relative change in } X} \]
- In a Log–Lin model multiply $\beta_2$ by 100 for a more meaningful figure.
- In a Lin-Log model divide $\beta_2$ by 100 for a more meaningful figure.

Reciprocal Models

- As $X$ increases indefinitely, $Y$ approaches the asymptotic value $\beta_1$.
  \[ Y_t = \beta_1 + \beta_2 \left( \frac{1}{X_t} \right) + u_t \]
- Example:
  - As per capita GNP increases, initially there is dramatic drop in CM (child mortality) but the drop tapers off as per capita GNP continues to increase.
  \[ \bar{CM}_t = 81.79436 + 27,273.17 \left( \frac{1}{\text{PGNP}} \right) \]

Logarithmic Reciprocal Model

- Log hyperbola or logarithmic reciprocal model takes the form:
  \[ \ln Y_t = \beta_1 - \beta_2 \left( \frac{1}{X_t} \right) + u_t \]
  \[ \frac{dY}{dX} = \beta_2 \frac{Y}{X^2} \]
  - Initially $Y$ increases at an increasing rate and then increases at a decreasing rate.

Functional Forms

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
<th>Slope $\left( \frac{dY}{dX} \right)$</th>
<th>Elasticity $\left( \frac{dY}{dX} \cdot \frac{X}{Y} \right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$Y = \beta_1 + \beta_2 X$</td>
<td>$\beta_2$</td>
<td>$\beta_2 \left( \frac{X}{Y} \right)^*$</td>
</tr>
<tr>
<td>Log-linear</td>
<td>$\ln Y = \beta_1 + \beta_2 \ln X$</td>
<td>$\beta_2 \left( \frac{Y}{X} \right)$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>Log-lin</td>
<td>$\ln Y = \beta_1 + \beta_2 X$</td>
<td>$\beta_2 \left( \frac{Y}{X} \right)$</td>
<td>$\beta_2 \left( \frac{X}{Y} \right)^*$</td>
</tr>
<tr>
<td>Lin-log</td>
<td>$Y = \beta_1 + \beta_2 \ln X$</td>
<td>$\beta_2 \left( \frac{Y}{X} \right)$</td>
<td>$\beta_2 \left( \frac{1}{X} \right)^*$</td>
</tr>
<tr>
<td>Reciprocal</td>
<td>$Y = \beta_1 + \beta_2 \left( \frac{1}{X} \right)$</td>
<td>$-\beta_2 \left( \frac{1}{X^2} \right)$</td>
<td>$-\beta_2 \left( \frac{1}{X^2} \right)^*$</td>
</tr>
<tr>
<td>Log reciprocal</td>
<td>$\ln Y = \beta_1 - \beta_2 \left( \frac{1}{X} \right)$</td>
<td>$\beta_2 \left( \frac{Y}{X^2} \right)$</td>
<td>$\beta_2 \left( \frac{1}{X} \right)^*$</td>
</tr>
</tbody>
</table>
Choice of Functional Form

- Choice of a functional form is easy in bivariate cases, because we can plot the variables and get some rough idea.
- The choice becomes much harder in multivariate cases involving more than one regressor.
- A great deal of skill and experience are required in choosing an appropriate model for empirical estimation.

Tips

- Keep in mind:
  - The underlying theory may suggest a particular functional form.
  - The knowledge of slope and elasticity coefficients of the various models will help to compare the various models.
  - The coefficients of the model chosen should satisfy certain a priori expectations.
  - Sometime more than one model may fit a given set of data reasonably well.
  - Do not overemphasize the $r^2$ measure in the sense that the higher the $r^2$ the better the model.

Error Term in Transformation

- Consider this regression model: $Y_i = \beta_1 X_i^{\beta_2}$
- Error term may enter in three forms:
  
  $Y_i = \beta_1 X_i^{\beta_2} u_i$
  $\ln Y_i = \alpha + \beta_2 \ln X_i + \ln u_i$
  
  $Y_i = \beta_1 X_i^{\beta_2} e^{u_i}$
  $\ln Y_i = \alpha + \beta_2 \ln X_i + u_i$
  
  $Y_i - \beta_1 X_i^{\beta_2} + \epsilon_i$
  $\ln Y_i = \ln (\beta_1 X_i^{\beta_2}) + \epsilon_i$

- 1st and 2nd models are intrinsically linear-regression models and the 3rd is intrinsically nonlinear-in-parameter.

Error Term in Transformation

- Although following linear regression models can be estimated by OLS or ML, we have to be careful about the properties of the stochastic error term that enters these models.
  
  $\ln Y_i = \alpha + \beta_2 \ln X_i + \ln u_i$
  
  $\ln Y_i = \alpha + \beta_2 \ln X_i + u_i$

  - BLUE property of OLS requires that $u_i$ has zero mean value, constant variance, and zero autocorrelation.
  - For hypothesis testing, we further assume that $u_i \sim N(0, \sigma^2)$.
  - Thus in the 1st model we must test the normality on $\ln u_i$
Homework 3A

Basic Econometrics (Gujarati, 2003)

1. Chapter 6, Problem 13 [50 points]
2. Chapter 6, Problem 14 [50 points]

Assignment weight factor - 0.5