Applied Econometrics for Researchers

Ordinary Least Squares (Cont.)

H.C. Kongsted

Department of Strategy and Innovation
Copenhagen Business School
Denmark

Outline of Lecture

The Regression Model

- Recall: Regression is a method used to quantify any linear association between–on the one side–a dependent variable and–on the other side–one or more independent variables
- In case we have multiple x variables, the regression is written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i \tag{1}$$

  k variables, one coefficient for each variable, and a constant term.
- In Workshop 1, you worked with a multiple regression in Stata:

```
reg prodnew extsource rdintpct inconst lempl00
```

Why Multiple Regression?

- Why don't we just stick to simple correlation and t-tests instead of doing (multiple) regression analysis?
  - Correlations and simple bivariate models are likely to suffer from what we call spurious correlation since relevant factors are not controlled for ("correlated omitted variables")
- We want to ask the question "what is the effect of a unit change in x, *keeping everything else constant?*"
- But before we do that, let's assess the validity and usefulness of our model.

Goodness of Fit I

- We would often like to know how well the model describes the data in general
- For this we use measures of the overall "goodness of fit" of the model
- We obtain what is called the Total Sum of Squares which measures the deviation of the actual values of the dependent variable from their overall average

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2}$$

Goodness of Fit II

Total sum of squares may be decomposed into two parts

- First, the *explained* part is the distance between the predicted values and the mean value

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3}$$

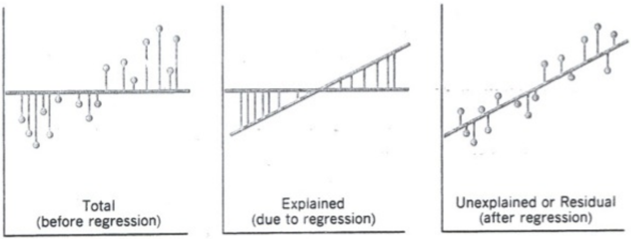
- Second, the *unexplained* or *residual* part is the distance between the predicted values and the observed values also known as the Residual Sum of Squares

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i{}^2 \tag{4}$$

where $\hat{u}_i$ is the residual, the "left-over" part of the dependent variable for observation i.

Overall Goodness of Fit

- SST= SSE + SSR



Goodness of Fit Statistics

- We can now evaluate how much of the total variation that is in fact explained by looking at the ratio between the explained compared to the total

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \qquad (5)$$

- $R^2 = 1$: All observations are located on the regression line - perfectly explained
- $R^2 = 0$: No explanatory power at all
- R^2 is a measure that indicates the *share* of the total variation explained by the model

Adjusted Goodness of Fit

- Note: if you add another regressor to the model, R^2 will increase (or remain unchanged). It never falls, hence a large R^2 may result simply from adding (potentially irrelevant) regressors to the model
- Alternatively we may use the *adjusted* R^2 where k is the number of explanatory variables and n the number of observations:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1))} \tag{6}$$

Overall Significance of the Model

- Worth understanding if the model *as a whole* has some relevance - whether we can rule out that *all* slope parameters are equal to zero (then the model would explain nothing)
- We do that by using the so-called F-test with
  $H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$ against
  H_a : At least one β coefficient is non-zero

$$F = \frac{\text{Explained Variance}}{\text{Unexplained Variance}} \quad (7)$$

$$= \frac{SSE/k}{SSR/(n-k-1)} = \frac{(SST - SSR)/k}{SSR/(n-k-1)} \quad (8)$$

The Model Yard-Stick

- The F test indicates the overall relevance of the model:
- We use the calculated F-value and find a corresponding probability (p-value) that all the estimated slope parameters are equal to zero - this is the significance of the model
  - The "degrees of freedom" of the numerator is the number of explanatory variables
  - The "degrees of freedom" of the denominator is number of observations minus number of explanatory variables minus 1
- Note: The F-statistic for one particular hypothesis - just one of many different F-tests that we will need during this course! Workshop 2 looks into a test of excluding a subset of variables from the model.
- In a bivariate model: F test is simply the squared t test

UK Innovation example

```
. reg prodnew extsource rdintpct inconst lempl00

      Source |       SS       df       MS              Number of obs =     431
-------------+------------------------------           F(4, 426)     =    8.40
       Model |  9180.92543     4  2295.23136           Prob > F      =  0.0000
    Residual |  116446.369   426  273.348285           R-squared     =  0.0731
-------------+------------------------------           Adj R-squared =  0.0644
       Total |  125627.295   430  292.156499           Root MSE      =  16.533
```

The Main Statistical Issues

- Our OLS procedure produces an estimate—it is *not* the unknown true value of β, just our best "guess."
- For e.g. $\hat{\beta}_1$ to be a "good" guess, it should satisfy three statistical properties:
  - $\hat{\beta}_1$ should be **centered** around its true (but unknown) value if we did "repeated sampling": Unbiased estimation
  - $\hat{\beta}_1$ should vary in a **narrow range** around its true (but unknown) value: Efficiency of estimation
  - $\hat{\beta}_1$ should become even more narrowly distributed around its true value if we somehow obtained a **larger sample**: Consistent estimation.
- We will focus our attention on consistency and efficiency.

The OLS Assumptions

1. The regression model is linear in β and u
2. The observations have been obtained as a *random* sample
3. No x variable is a linear function of (one or more) of the other x's (no exact multicollinearity)
4. u has a zero population mean and all x's are uncorrelated with u (zero correlation)
5. u has a constant variance (no heteroskedasticity)
6. (u is normally distributed)

Assumptions 1.-4. implies consistency. Assumptions 1.-5. implies efficiency. Assumption 6 is optional and not needed in "large" samples.

Multicollinearity

- If two regressors are highly correlated: it is difficult to distinguish the effect of one variable from the other: Ex. We control for firm size by including sales and number of employees.
- It is a model design problem.
- Consequences:
  1. Variance of estimates increase and t-scores fall
  2. Estimates changes substantially by minor specification alterations, but goodness-of-fit statistics are largely unchanges if a variable is added
  3. Estimates of non-collinear variables can be largely unaffected

Detecting Multicollinearity

Detecting:

- High correlation coefficients between independent variables
- High R^2 with all low t-scores
- Look at the Variance Inflation Factor (VIF) – calculated for each variable as $\frac{1}{1-R_j^2}$ where R_j^2 is the share of variance of the j^{th} explanatory variable that can be explained by the *other* explanatory variables in an auxiliary regression – $VIF > 10$ suggest multicollinearity

Remedy:

- Drop one of more of the collinear variables?
- Transform the collinear variables?
- Increase sample size (?)

Homoskedasticity versus Heteroskedasticity

- Another assumption of OLS is that the variance of the error term is constant across the observations
- We are faced with two possibilities:

$$Var(u_i) = \sigma^2 \qquad (i = 1, 2, 3, \ldots, n) \qquad (9)$$
$$Var(u_i) = \sigma_i^2 \qquad (i = 1, 2, 3, \ldots, n) \qquad (10)$$

- Where (9) is a case of homoskedastic error terms and (10) is a case of heteroskedastic error terms

Efficiency

Why is potential heteroskedasticity of interest?

- We say that an estimator is efficient if it is able to produce estimates with low variance
  - Relatively high variance indicates inefficient estimator
  - Relatively low variance indicates efficient estimator
- Assumptions 1.-5. imply that OLS is in fact the best ("most efficient") of all estimators that are linear and unbiased.
- Main concern: Usual OLS standard errors are not valid if there is heteroskedasticity.

Detecting and remedying heteroskedasticity

Detection:

- Several tests are available: The residuals $\hat{u}_i$ are key to this.
- In Stata: Look through the post estimation commands - particularly `hettest` and `imtest`
- Two of the most used tests are the Breusch-Pagan test and the White test: Run auxiliary regression of $\hat{u}^2$ on independent variables, squares, cross-products.

Main remedy:

- Run the regression with Huber-White sandwich corrected standard errors – `robust` option in Stata: Produces valid standard errors for OLS estimates even under heteroskedasticity.

Look into some specification issues

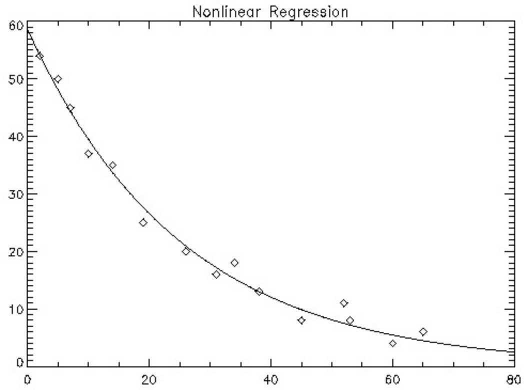
How can we make the linear regression model applicable to a broader range of issues and hypotheses?

- Dummy variables: Qualitative information in the regression model (see separate slides)
- Interactions: Letting the effect of one variable be dependent on the value of another variable (see separate slides)
- Introduce non-linearities in the linear regression model

Linear versus non-linear

- The linear regression models a relationship as additive - and thus equations as linear ($y_i = \beta_0 + \boldsymbol{\beta'} \mathbf{x}_i + u_i$) - straight lines with constant slopes and an additive error term
- Sometimes we would like to estimate non-linear functional forms – the relationship between two variables is not a straight line but (perhaps) "curvilinear"
- Important implications: The marginal change in y as x increases will be different as we move between different points along the x-axis
- After respecification, use OLS to allow for a wide range of different functional relationships

Example: Y decreases with X but a decreasing rate



Useful transformations

- Functional forms that are often used and may be transformed into linear functions by simple tricks

Function	Functional Equation	Linear Regression Equation	Dependent Variable	Independent Variable
Exponential (semi-log)	$y = \beta_0 e^{\beta_1 x}$	$\log(y) = \log \beta_0 + \beta_1 x$	$\log(y)$	x
Power (log-log)	$y = \beta_0 x^{\beta_1}$	$\log(y) = \log(\beta_0) + \beta_1 \log(x)$	$\log(y)$	$\log(x)$
Polynomial (2^{nd}-degree)	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	As function	y	x, x^2

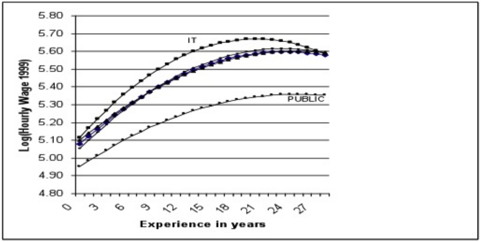
Which Function to Choose?

- Does theory suggest eg a diminishing marginal effect? Or increasing?
- Observe existing empirical studies (conventions)
- Know the properties of candidate functional forms: Stata:
  - `twoway function y=1+4*x-0.1*x**2, range(0 6)`
- Let the data speak? Start from a general starting point and reduce the model by hypothesis testing. Issues?

Example: Non-linear Wage Equation

Wage profiles for engineers across experience. Main industries, 1999.

Male construction engineer with diploma-degree [reference group]



Exponential and polynomial:

$$\ln(w) = b_1 + b_2 EXP + b_3 EXP^2 + b_4 EXP{*}IND + b_5 EXP^2{*}IND$$
$$+ \text{Dummies (IND (industry), gender, education)}$$

Interpretation of Functional Forms:

In some cases, the interpretation of the regression coefficients is fairly straightforward:

- Semi-log: 100 times β_1 is approximately the percentage increase in y with a unit increase in x
- Power (log-log): β_1 is the elasticity of y with respect to x

The parameter estimate of interest is given directly from the transformed regression.
Other cases need some computation:

- Differentiate the terms that involve x and evaluate at given $x-$values (Stata: `margins`)

Careful with extending the relationship beyond the range of x observed in the data!

What's next?

- Friday: Workshop 2 - hypothesis testing, interactions; further analysis of the UK CIS data set.
- Via Zoom.
- Next week: Back on a regular schedule: Lectures November 2/Workshop November 4.