Linear Regression

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Linear Regression Overview

- Linear regression examples
- Linear regression model
- Estimated regression line
- Single versus multiple regression
- Coefficients and marginal effects
- Goodness of fit (R-squared)
- Hypothesis testing for coefficient significance
  - t-test for a single coefficient significance
  - F-test for multiple coefficients significance

Linear Regression

Linear regression examples

- Explain student grades using the number of hours studied
- Explain the effect of education on income
- Explain the effect of the number of bedrooms on house prices
- Explain the effect of the recession on stock prices

Linear regression set up

- Regression analysis does not establish a cause-and-effect relationship, just that there is a relationship.
- The cause-and-effect relationship must be determined using a theoretical model or a logical reason.
- The dependent variable is a continuous variable.
- The independent variables can take any form - continuous or discrete or indicator variables.
- The simple linear regression model has one independent variable.
- The multiple linear regression model has two or more independent variables.

Linear regression model

Linear regression model

- The linear regression model describes how the dependent variable is related to the independent variable(s) and the error term:

$$y = \beta_0 + \beta_1 x_1 + u$$

or

$$y = x'\beta + u$$

- y is the dependent variable (explained, predicted, or response variable)
- x is the independent variables (control variables or regressors)
- β are unknown parameters to be estimated
    - β_0 is the intercept
    - β_1 is the slope
- u is the error term or disturbance

Estimated regression equation

- The estimated regression equation shows how to calculate predicted values of the dependent variable using the values of the independent variable(s).

$$\hat{y} = b_0 + b_1 x_1 = x'b$$

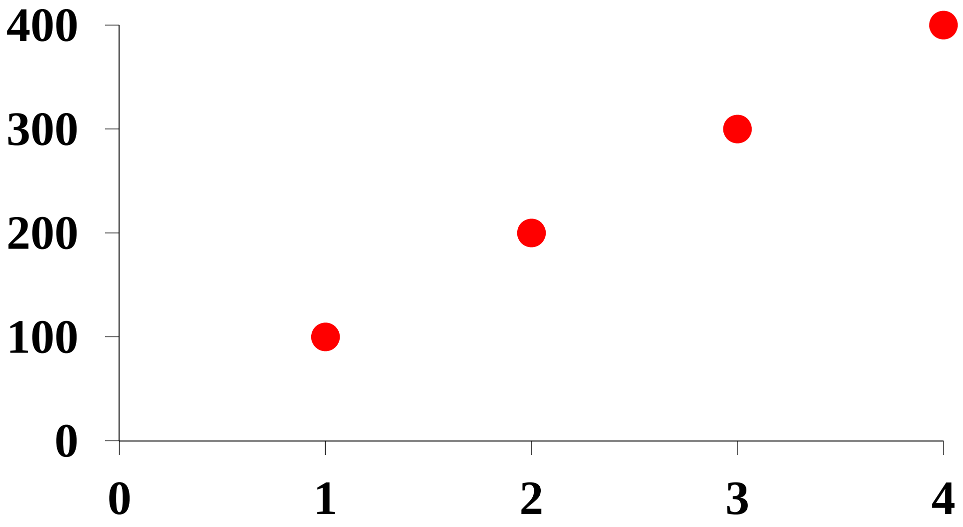
- Interpretation of the coefficients: one unit increase in *x* will increase the dependent variable *y* by b_1 units.
- Note that there is no error term when we predict the value of the depended variable.

- Regression residuals are calculated as the difference between the actual and predicted values of the dependent variable:

$$u = y - \hat{y} = y - b_0 - b_1 x_1 = y - x'b$$

Simple linear regression examples

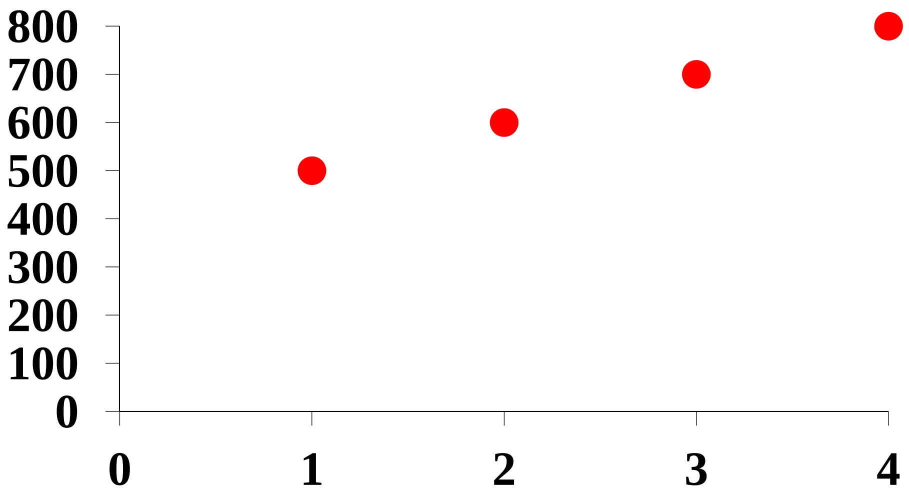
Regression line



x = number of credit cards
y = dollars spent
For each additional credit card, a person spends $100 more.

The equation for the line is $\hat{y} = b_0 + b_1 x_1 = 0 + 100x_1$
intercept = b_0 = 0 (when x_1=0, then $\hat{y} = b_0$)
slope = b_1 = 100 (when x_1 increases by 1, then $\hat{y}$ increases by b_1)

Regression line, new example with a positive intercept



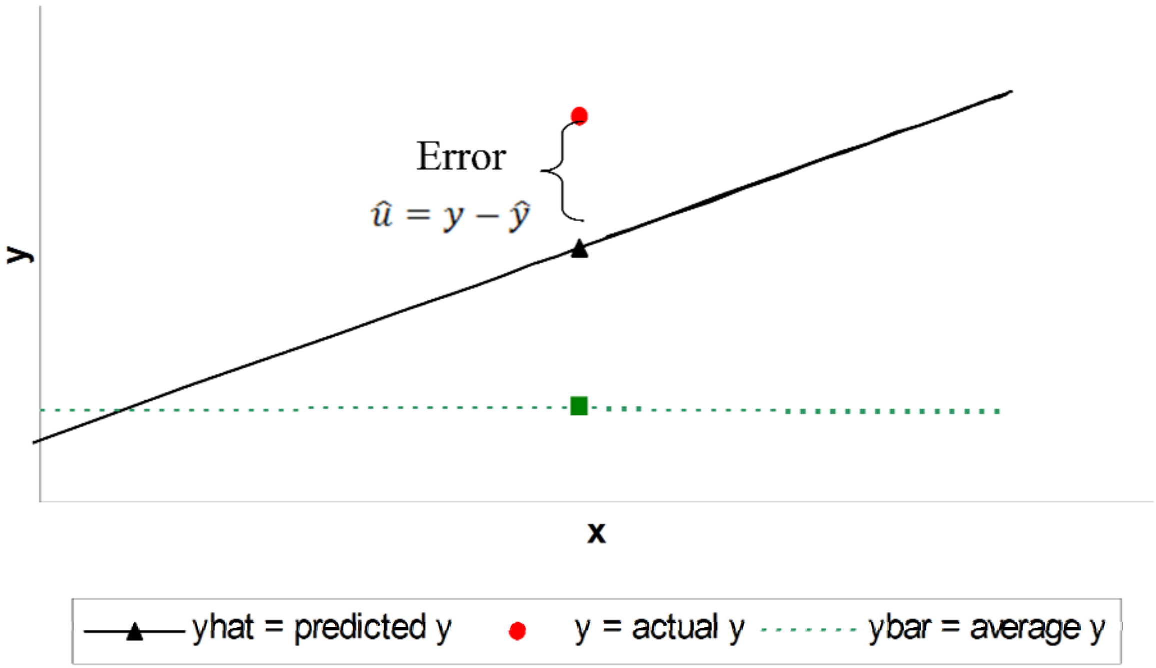
The equation for the line is $\hat{y} = b_0 + b_1 x_1 = 400 + 100x_1$

intercept = b_0 = 400 (when x_1=0, then $\hat{y} = b_0$)

slope = b_1 = 100 (when x_1 increases by 1, then $\hat{y}$ increases by b_1)

For each additional credit card, a person spends $100 more.

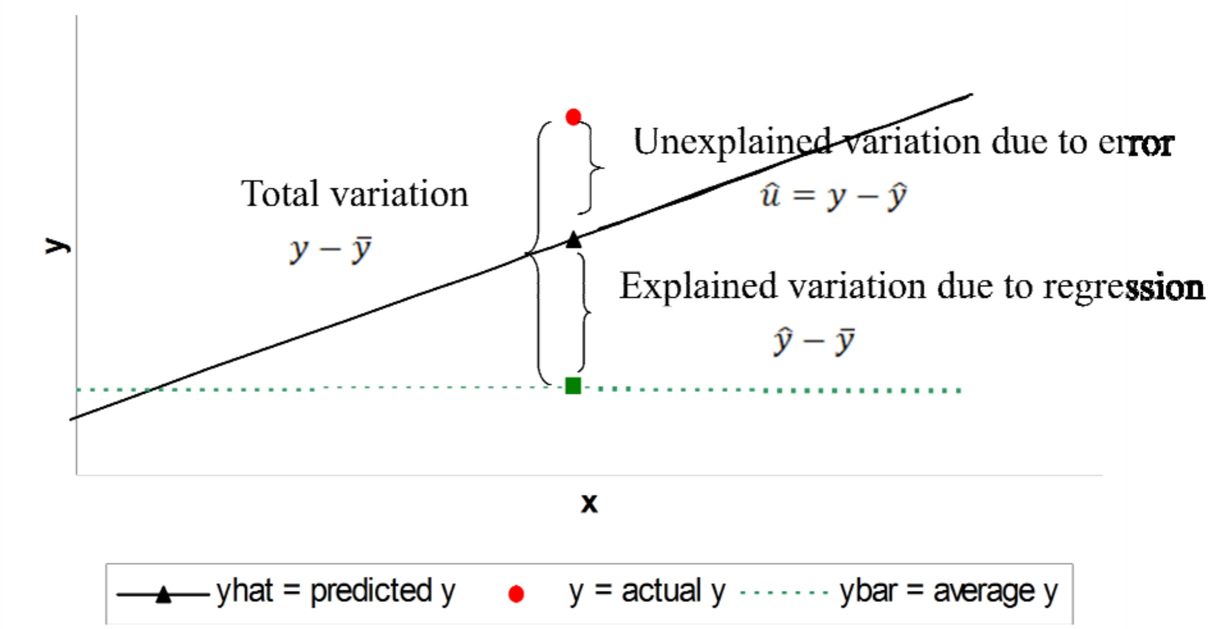
Regression error



The error is the difference between the actual values and the predicted values of the dependent variable:

$$u = y - \hat{y} = y - b_0 - b_1 x_1$$

Variations: total variation, explained variation and unexplained variation



$$\sum(y-\bar{y})^2=\sum(\hat{y}-\bar{y})^2+\sum(y-\hat{y})^2$$

Total variation = explained variation due to regression + unexplained variation due to error
sum of squares total = sum of squares due to regression + sum of squares due to error
SST= SSR+SSE

The least squares method (OLS: ordinary least squares)

- The least squares method is used to calculate the coefficients so that the errors are as small as possible.
- We minimize the sum of squared residuals:

$$\sum u^2 = \sum (y - \hat{y})^2 = \sum (y - b_0 - b_1 x)^2$$

- In a simple linear regression the coefficients are calculated as:

$$b_1 = \frac{cov(x, y)}{var(x)}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

OLS regression in matrix form

- The regression line is specified as:

$$E(y|x) = x'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_p$$

- Marginal effects in the linear regression model are the coefficients.

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j$$

- In multiple linear regression, the coefficients are calculated as:

$$b = \left(x'x\right)^{-1}(x'y)$$

- Assumptions of the OLS estimator:
  - Exogeneity of regressors
  - Homoscedasticity
  - Uncorrelated observations

Goodness of fit

R-squared

- The coefficient of determination (R-squared or R^2) provides a measure of the goodness of fit for the estimated regression equation.
- R^2 = SSR/SST = 1 – SSE/SST
- Values of R^2 close to 1 indicate perfect fit, values close to zero indicate poor fit.
- R^2 that is greater than 0.25 is considered good in the economics field.
- R-squared interpretation: if R-squared=0.8 then 80% of the variation is explained by the regression and the rest is due to error. So, we have a good fit.

Adjusted R-squared

- Problem: R^2 always increases when a new independent variable is added. This is because the SST is still the same but the SSE declines and SSR increases.
- Adjusted R-squared corrects for the number of independent variables and is preferred to R-squared.

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- where p is the number of independent variables, and n is the number of observations.

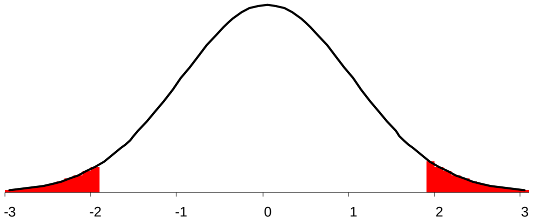
t-test for significance of one coefficient

- The t-test is used to determine whether the relationship between y and x_j is significant.

$$H_0\colon\ \beta_j = 0 \qquad\qquad H_a\colon\ \beta_j \neq 0$$

- The null hypothesis is that the coefficient is not significantly different than zero.
- The alternative hypothesis is that the coefficient is significantly different from zero.

- We use the t-distribution:
  - The test statistic t = coefficient /standard error
  - The critical values are from the t distribution
  - The test is a two-tailed test.



- Reject the null hypothesis and conclude that coefficient is significantly different from zero if:
  - The test statistic t is in the critical rejection zone
  - The p-value is less than 0.05
- The goal is to find coefficients that are significant.

F-test for overall significance of all coefficients

- Testing whether the relationship between y and all x variables is significant.
- The null hypothesis is that the coefficients are not jointly significantly different from zero.
- The alternative hypothesis is that the coefficients are jointly significantly different from zero.

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a\colon \beta_1 \neq 0\ or\ \beta_2 \neq 0\ or\ \ldots\ \beta_p \neq 0$$

- Use the F-distribution
  - The test statistic F = MSR/MSE
  - The critical values are from the F distribution
  - The F-test is an upper one-tail test

ANOVA table

Total variation = explained variation due to regression + unexplained variation due to error

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic
Regression	SSR=$\sum(\hat{y} - \bar{y})^2$	p = number of independent variables	MSR=SSR/p	F=MSR/MSE
Error	SSE=$\sum(y - \hat{y})^2$	n-p-1	MSE=SSE/(n-p-1)	
Total	SST=$\sum(y - \bar{y})^2$	n-1 (n=number of observations)		

- Find critical values in the F table (significance level =0.05)
    - degrees of freedom in the numerator = number of independent variables = p
    - degrees of freedom in the denominator = n-p-1

- Reject the null hypothesis if the F-test statistic is greater than the F-critical value.
- Reject the null hypothesis if the p-value is less than 0.05.
- The goal is to find a regression model with coefficients that are jointly significant.

Panel Data Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Panel Data Models Overview

- Panel data characteristics, panel data types
- Variation types (overall, within, and between variation)
- Panel data models (pooled model, fixed effects model, and random effects model)
- Estimator properties (consistency and efficiency)
- Estimators (pooled OLS, between, fixed effects, first differences, random effects)
- Tests for choosing between models (Breusch-Pagan LM test, Hausman test)

Panel Data Models

Panel data model examples

- Labor economics: effect of education on income, with data across time and individuals.
- Economics: effects of income on savings, with data across years and countries.

Panel data characteristics

- Panel data provide information on individual behavior, both across individuals and over time – they have both cross-sectional and time-series dimensions.
- Panel data include N individuals observed at T regular time periods.
- Panel data can be balanced when all individuals are observed in all time periods ($T_i = T$ for all i) or unbalanced when individuals are not observed in all time periods ($T_i \neq T$).
- We assume correlation (clustering) over time for a given individual, with independence over individuals.
  - Example: the income for the same individual is correlated over time but it is independent across individuals.

Panel data types

- Short panel: many individuals and few time periods (we use this case in class)
- Long panel: many time periods and few individuals
- Both: many time periods and many individuals

Regressors

- Varying regressors x_{it}.
  - annual income for a person, annual consumption of a product
- Time-invariant regressors $x_{it} = x_i$ for all t.
  - gender, race, education
- Individual-invariant regressors $x_{it} = x_t$ for all i.
  - time trend, economy trends such as unemployment rate

Variation for the dependent variable and regressors

- Overall variation: variation over time and individuals.
- Between variation: variation between individuals.
- Within variation: variation within individuals (over time).

Id	Time	Variable	Individual mean	Overall mean	Overall deviation	Between deviation	Within deviation	Within deviation (modified)
i	t	x_{it}	$\bar{x}_i$	$\bar{x}$	$x_{it} - \bar{x}$	$\bar{x}_i - \bar{x}$	$x_{it} - \bar{x}_i$	$x_{it} - \bar{x}_i + \bar{x}$
1	1	9	10	20	-11	-10	-1	19
1	2	10	10	20	-10	-10	0	20
1	3	11	10	20	-9	-10	1	21
2	1	20	20	20	0	0	0	20
2	2	20	20	20	0	0	0	20
2	3	20	20	20	0	0	0	20
3	1	25	30	20	5	10	-5	15
3	2	30	30	20	10	10	0	20
3	3	35	30	20	15	10	5	25

Individual mean $\bar{x}_i = \frac{1}{T} \sum_t x_{it}$

Overall mean $\bar{x} = \frac{1}{NT} \sum_i \sum_t x_{it}$

Overall variance $s_O^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \bar{x})^2$

Between variance $s_B^2 = \frac{1}{N-1} \sum_i (\bar{x}_i - \bar{x})^2$

Within variance $s_W^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \bar{x}_i)^2 = \frac{1}{NT-1} \sum_i \sum_t (x_{it} - \bar{x}_i + \bar{x})^2$

The overall variation can be decomposed into between variation and within variation.

$$s_O^2 \approx s_B^2 + s_W^2$$

- Time-invariant regressors (race, gender, education) have zero within variation.
- Individual-invariant regressors (time, economy trends) have zero between variation.
- We need to check the data to see if the between or within variation is larger for each variable.

Panel data models

- Panel data models describe the individual behavior both across time and across individuals. There are three types of models: the pooled model, the fixed effects model, and the random effects model.

Pooled model

- The pooled model specifies constant coefficients, the usual assumptions for cross-sectional analysis.

$$y_{it} = \alpha + \mathbf{x}_{it}'\beta + u_{it}$$

- This is the most restrictive panel data model and is not used much in the literature.

Individual-specific effects model

- We assume that there is unobserved heterogeneity across individuals captured by α_i.
    - Example: unobserved ability of an individual that affects wages.
- The main question is whether the individual-specific effects α_i are correlated with the regressors. If they are correlated, we have the fixed effects model. If they are not correlated, we have the random effects model.

Fixed effects model (FE)

- The FE model allows the individual-specific effects α_i to be correlated with the regressors x.
- We include α_i as intercepts.
- Each individual has a different intercept term and the same slope parameters.

$$y_{it} = \alpha_i + \mathbf{x}_{it}^{'}\beta + u_{it}$$

- We can recover the individual specific effects after estimation as:

$$\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i^{'}\hat{\beta}$$

  In other words, the individual-specific effects are the leftover variation in the dependent variable that cannot be explained by the regressors.
- Time dummies can be included in the regressors x.

Random effects model (RE)

- The RE model assumes that the individual-specific effects α_i are distributed independently of the regressors.
- We include α_i in the error term.
  Each individual has the same slope parameters and a composite error term $\varepsilon_{it} = \alpha_i + e_{it}$.

$$y_{it} = \mathbf{x}_{it}^{'}\beta + (\alpha_i + e_{it})$$

  Here $var(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_e^2$ and $cov(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2$
  so $\rho_\varepsilon = cor(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_e^2)$
- Rho is the interclass correlation of the error. Rho is the fraction of the variance in the error due to the individual-specific effects. It approaches 1 if the individual effects dominate the idiosyncratic error.

Panel data estimators

- The panel data models can be estimated with several estimators.
- The estimators differ based on whether they consider the between or within variation in the data.
- Their properties (consistency) differ based on which model is appropriate.

Estimator properties

- We prefer estimators that are consistent and efficient. We check for consistency first and then for efficiency.

Consistency

- The distribution of $\hat{\beta}_n$ collapses on β as n becomes large:

$$plim\, \hat{\beta}_n = \beta$$

- Consistency is established based on the law of large numbers.
- If an estimator is consistent, more observations will tend to provide more precise and accurate estimates.

Efficiency

- Efficiency (minimum variance) is usually established relative to specific classes of estimators.
  - Example: OLS is efficient (minimum variance) among the class of linear, unbiased estimators (Gauss-Markov Theorem).
  - Maximum likelihood (given correct distributional assumptions) is asymptotically efficient among consistent estimators.

Pooled OLS estimator

- The pooled OLS estimator uses both the between and within variation to estimate the parameters.
- The pooled OLS estimator is obtained by stacking the data over *i* and *t* into one long regression with *NT* observations and estimating it by OLS:

$$y_{it} = \alpha + \mathbf{x}_{it}^{'}\beta + (\alpha_i - \alpha + e_{it})$$

- If the true model is the pooled model and the regressors are uncorrelated with the error terms, the pooled OLS regressor is consistent.
- If the true model is fixed effects then the pooled OLS regressor is inconsistent.
- We need to have panel-corrected standard errors.

Between estimator

- The between estimator only uses the between variation (across individuals).
- It uses the time averages of all variables.
    - If an individual has a work experience of 9, 10, and 11 years measured over 3 periods then the average experience is 10.
- This is an OLS estimation of the time-averaged dependent variable on the time-averaged regressors for each individual.

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\beta + (\alpha_i - \alpha + \bar{e}_i)$$

- The number of observations is *N*. The time variation is not considered and the data are collapsed with one observation per individual.
- This estimator is seldom used because the pooled and RE estimators are more efficient.

Within estimator or fixed effects estimator

- The within estimator uses the within variation (over time).
- It uses time-demeaned variables (the individual-specific deviations of variables from their time-averaged values).

    - If an individual has a work experience of 9, 10, an 11 years measured over 3 periods, the average experience is 10. So the time-demeaned values are -1, 0, and 1.
- This is an OLS estimation of the time-demeaned dependent variable on the time-demeaned regressors.

$$y_{it} - \overline{y_\imath} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\beta + (e_{it} - \bar{e}_i)$$

Some software packages estimate:

$$y_{it} - \overline{y_\imath} + \bar{y} = \alpha + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i + \bar{\mathbf{x}})'\beta + (e_{it} - \bar{e}_i + \bar{e})$$

- The number of observations is *NT*.
- The individual-specific effects α_i cancel out.
- Here, α is the average of the individual effects.
- A limitation of the within estimator is that time-invariant variables are dropped from the model and their coefficients are not identified.
    - A female/male will have values of 1/0 for the female dummy variable, so the values minus the mean values (calculated over time) for each individual will be zero.
    - If we are interested in the effects of time-invariant variables, we need to consider different models (OLS or between estimators).

First-differences estimator

- The first-difference estimator uses the one-period changes for each individual.
- It uses first-differenced variables (the individual-specific one-period changes for each individual).
  - If an individual has a work experience of 9, 10, and 11 years measured over 3 periods then the first difference experience are missing (.), 1, and 1.
- This is an OLS estimation of the one-period changes of the dependent variable on the one-period changes in the regressors.

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (e_{it} - e_{i,t-1})$$

- The number of observations is *N*(*T*-1). We lose the first observation for each individual because of differencing.
- The individual-specific effects α_i cancel out.
- A limitation of the first-differences model is that time-invariant variables are dropped from the model and their coefficients are not identified.

Random effects estimator

- This is an OLS estimation of the transformed model:

$$y_{it} - \hat{\lambda}\overline{y_\imath} = \left(1 - \hat{\lambda}\right)\mu + (\mathbf{x}_{it} - \hat{\lambda}\bar{\mathbf{x}}_i)'\beta + \nu_{it}$$

$$\nu_{it} = \left(1 - \hat{\lambda}\right)\alpha_i + (e_{it} - \hat{\lambda}\bar{e}_i)$$

$$\lambda = 1 - \sigma_e \Big/ \sqrt{\sigma_e^2 + \sigma_\alpha^2}$$

- The number of observations is *NT*.
- The individual-specific effects α_i are in the error term.
- Note that $\hat{\lambda} = 0$ corresponds to pooled OLS and $\hat{\lambda} = 1$ corresponds to the within (fixed effects) estimator.
- The random effects estimates are a weighted average of the between and within estimates.
- The random effects estimator is fully efficient under the random effects model.

Models and estimators

Estimator/true model	Pooled model	Random effects model	Fixed effects model
Pooled OLS estimator	Consistent	Consistent	Inconsistent
Between estimator	Consistent	Consistent	Inconsistent
Within or fixed effects estimator	Consistent	Consistent	Consistent
First differences estimator	Consistent	Consistent	Consistent
Random effects estimator	Consistent	Consistent	Inconsistent

- The fixed effects estimator will always give consistent estimates, but they may not be the most efficient.
- The random effects estimator is inconsistent if the appropriate model is the fixed effects model.
- The random effects estimator is consistent and most efficient if the appropriate model is random effects model.

Choosing between fixed and random effects

Breusch-Pagan Lagrange Multiplier test

- This is a test for the random effects model based on the OLS residual.
- Test whether σ_u^2 or equivalently $cor(u_{it}, u_{is})$ is significantly different from zero.
- If the LM test is significant, use the random effects model instead of the OLS model.
- We still need to test for fixed versus random effects.

Hausman test

- The random effects estimator is more efficient so we need to use it if the Hausman test supports it. If it does not support it, use the fixed effects model.
- Hausman test tests whether there is a significant difference between the fixed and random effects estimators.
- The Hausman test statistic can be calculated only for the time-varying regressors.
- The Hausman test statistics is:

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE}))(\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

- It is chi-square distributed with degrees of freedom equal to the number of parameters for the time-varying regressors.
- If the Hausman test is insignificant use the random effects.
- If the Hausman test is significant use the fixed effects.

Probit and Logit Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Probit and Logit Models Overview

- Examples of probit and logit models
- Binary dependent variable
- Linear regression model, probit, and logit models functional forms and properties
- Model coefficients and interpretations
- Marginal effects (and odds ratios) and interpretations
- Goodness of fit statistics (percent correctly predicted and pseudo R-squared)
- Choice between probit and logit
- Economic models that lead to use of probit and logit models

Probit and Logit Models (Binary Outcome Models)

Binary outcome examples

- Consumer economics: whether a consumer makes a purchase or not.
- Labor economics: whether an individual participates in the labor market or not.
- Agricultural economics: whether or not a farmer adopts or uses organic practices, marketing/production contracts, etc.

Binary outcome dependent variable

- The decision/choice is whether or not to have, do, use, or adopt.
- The dependent variable is a binary response
- It takes on two values: 0 and 1.

$$y = \begin{cases} 0 \; if \; no \\ 1 \; if \; yes \end{cases}$$

Binary outcome models

- Binary outcome models are among the most used in applied economics.
- A look at the OLS model: $y = \mathbf{x}'\beta + e$
- Binary outcome models estimate the probability that *y*=1 as a function of the independent variables.

$$p = \mathrm{pr}[y = 1|\mathbf{x}] = \mathrm{F}(\mathbf{x}'\beta)$$

- There are three different models depending on the functional form of $\mathrm{F}(\mathbf{x}'\beta)$.

Regression model (linear probability model)

- In the linear probability model, $\mathrm{F}(\mathrm{x}'\beta) = \mathrm{x}'\beta$

$$p = \mathrm{pr}[y = 1|\mathrm{x}] = \mathrm{x}'\beta$$

- A problem with the regression model is that the predicted probabilities will not be limited between 0 and 1.
- We do not use the regression model with binary outcome data.

Logit model

- For the logit model, $\mathrm{F}(\mathrm{x}'\beta)$ is the cdf of the logistic distribution.

$$\mathrm{F}(\mathbf{x}'\beta) = \Lambda(\mathbf{x}'\beta) = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}} = \frac{\exp{(\mathbf{x}'\beta)}}{1 + \exp{(\mathbf{x}'\beta)}}$$

- The predicted probabilities are limited between 0 and 1.

Probit model

- For the probit model, $\mathrm{F}(\mathbf{x}'\beta)$ is the cdf of the standard normal distribution.

$$\mathrm{F}(\mathbf{x}'\beta) = \Phi(\mathbf{x}'\beta) = \int_{-\infty}^{\mathbf{x}'\beta} \phi(z) dz$$

- The predicted probabilities are limited between 0 and 1.

Model coefficients

- Probit and logit models are estimated using the maximum likelihood method.

Interpretation of coefficients

- An increase in x increases/decreases the likelihood that y=1 (makes that outcome more/less likely). In other words, an increase in x makes the outcome of 1 more or less likely.
- We interpret the *sign* of the coefficient but not the *magnitude*. The magnitude cannot be interpreted using the coefficient because different models have different scales of coefficients.

Comparison of coefficients

- Coefficients differ among models because of the functional form of the *F* function.

$$\beta_{logit} \simeq 4\beta_{OLS}$$

$$\beta_{probit} \simeq 2.5\beta_{OLS}$$

$$\beta_{logit} \simeq 1.6\beta_{probit}$$

- We should not compare the magnitude of the coefficients among different models.

Marginal effects

- When estimating probit and logit models, it is common to report the marginal effects after reporting the coefficients.
- The marginal effects reflect the change in the probability of *y*=1 given a 1 unit change in an independent variable x.

Marginal effects for the regression model

- For the OLS regression model, the marginal effects are the coefficients and they do not depend on x.

$$\partial p/\partial \mathrm{x}_j = \beta_j$$

- The index *j* refers to the *j*th independent variable.
- [When we use the index *i*, it refers to the *i*th observation.]

Marginal effects for the binary models (probit and logit)

- For the logit and probit models, the marginal effects are calculated as:

$$\partial p/\partial \mathrm{x}_j = \mathrm{F}'(\mathbf{x}'\beta)\beta_j$$

- The marginal effects depend on x, so we need to estimate the marginal effects at a specific value of x (typically the means).
- Coefficients and marginal effects have the same signs because $\mathrm{F}'(\mathbf{x}'\beta) > 0$.

Marginal effects for the logit model

$$\partial p/\partial \mathrm{x}_j = \Lambda(\mathbf{x}'\beta)[1 - \Lambda(\mathbf{x}'\beta)]\beta_j = \frac{e^{\mathbf{x}'\beta}}{\left(1 + e^{\mathbf{x}'\beta}\right)^2}\beta_j$$

Marginal effects for the probit model

$$\partial p/\partial \mathrm{x}_j = \phi(\mathbf{x}'\beta)\beta_j$$

Estimating marginal effects

Marginal effects at the mean

- The marginal effects are estimated for the average person in the sample $\bar{\mathbf{x}}$.

$$\partial p/\partial \mathrm{x}_j = \mathrm{F}'(\bar{\mathbf{x}}'\beta)\beta_j$$

- Most papers report marginal effects at the mean.
- A problem is that there may not be such a person in the sample.

Average marginal effects

- The marginal effects are estimated as the average of the individual marginal effects.

$$\partial p/\partial \mathrm{x}_j = \frac{\sum \mathrm{F}'(\mathbf{x}'\beta)}{\mathrm{n}}\ \beta_j$$

- This is a better approach of estimating marginal effects, but papers still use the previous approach.
- In practice, the two ways to estimate marginal effects produce almost identical results most of the time.

Partial effects for discrete variables

- Predict the probabilities for the two discrete values of a variable and take the difference:

$$F(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(k + 1)) - F(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(k))$$

Interpretation of marginal effects

- An increase in x increases (decreases) the probability that y=1 by the marginal effect expressed as a percent.
  - For dummy independent variables, the marginal effect is expressed in comparison to the base category (x=0).
  - For continuous independent variables, the marginal effect is expressed for a one-unit change in x.
- We interpret both the sign and the magnitude of the marginal effects.
- The probit and logit models produce almost identical marginal effects.

Odds ratio/relative risk for the logit model

- The odds ratio or relative risk is p/(1-p) and measures the probability that y=1 relative to the probability that y=0.

$$p = \frac{\exp\,(\mathbf{x}'\beta)}{1 + \exp\,(\mathbf{x}'\beta)}$$

$$\frac{p}{1-p} = \exp\,(\mathbf{x}'\beta)$$

$$ln\frac{p}{1-p} = \mathbf{x}'\beta$$

- An odds ratio of 2 means that the outcome y=1 is twice as likely as the outcome of y=0.
- Odds ratios are estimated with the logistic model.
- Reporting marginal effects instead of odds ratios is more popular in economics.

Predicted probabilities and goodness of fit measures

- After estimating the models, we can predict the probability that y=1 for each observation.

$$\hat{p} = \mathrm{pr}[y = 1|\mathbf{x}] = \mathrm{F}\left(\mathbf{x}'\hat{\beta}\right)$$

- For the regression model, the predicted probabilities are *not* limited between 0 and 1.
- For the logit and probit models, the predicted probabilities are limited between 0 and 1.
- The predicted probability indicate the likelihood of y=1. If the predicted probability is greater than 0.5 we can predict that y=1, otherwise y=0.

Goodness of fit measures

Percent correctly predicted values

- If the predicted probability is greater than 0.5 we can predict that y=1, otherwise y=0.
- We can create the following table:

	Actual y=1	Actual y=0
Predicted yhat=1	True	False
Predicted yhat=0	False	True

- We have four cases of 0/1: two of them are correct predictions and two of them are wrong predictions.
- The percent correctly predicted values are the proportion of true predictions to total predictions.

Pseudo R-squared (McFadden R-squared)

- The pseudo R-square is calculated as:

$$\text{R-squared} = 1 - L_{ur}/L_r$$

- It compares the unrestricted log-likelihood L_{ur} for the model we are estimating and the restricted log-likelihood L_r with only an intercept.
- If the independent variables have no explanatory power, the restricted model will be the same as unrestricted model and R-squared will be 0.

Discussion about binary outcome models

Choice between the logit and probit model

- The choice depends on the data generating process, which is unknown.
- The models produce almost identical results (different coefficients but similar marginal effects).
- The choice is up to you.

Coding of the dependent variable

- If we reverse the categories 0 and 1, the signs of the coefficients are reversed (positive become negative and vice versa) but the magnitudes are the same.

Latent variable models

- A latent variable is a variable that is incompletely observed y^*. Latent variables can be introduced into binary outcome models in two ways: index functions and random utility models.

Index function models

- The latent variable is an index of the unobserved propensity for the event to occur.
- Index models are used in two step models, which will be covered later in class.
  - o Example: We cannot observe how much people want to work, only if they work or not.

$$y = \begin{cases} 1\ if\ y *> 0 \\ 0\ if\ y *\leq 0 \end{cases}$$

Random utility models

- The latent variable is the difference in utilities if the event occurs or does not occur.
- They are often a result of individual choice.
  - o Example: a consumer chooses one product or another depending on which utility is higher.

$$U_0 = V_0 + e_0$$

$$U_1 = V_1 + e_1$$

$$p(y = 1) = p(U_1 > U_0)$$

Bivariate Probit and Logit Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Bivariate Probit and Logit Models Overview

- Bivariate probit and logit models equations
- Coefficients and marginal effects

Bivariate probit model

Bivariate outcome examples

- Individual decision whether to work or not *and* whether to have children or not.
- Farmer decision of whether to use marketing contracts or not *and* whether to use environmental contracts or not.
- The bivariate models estimates decisions that are interrelated as opposed to independent.

Bivariate probit model specification

- The bivariate probit model is a joint model for two binary outcomes.
- These outcomes may be correlated, with correlation ρ.
- If the correlation turns out insignificant, then we can estimate two separate probit models, otherwise we have to use a bivariate probit model.
- The unobserved latent variables are presented as:

$$\mathrm{y}_1^* = \mathbf{x}_1^{'}\beta_1 + e_1$$
$$\mathrm{y}_2^* = \mathbf{x}_2^{'}\beta_2 + e_2$$

- The bivariate probit model specifies the outcomes as:

$$y_1 = \begin{cases} 1 \; if \; \mathrm{y}_1^* > 0 \\ 0 \; if \; \mathrm{y}_1^* \leq 0 \end{cases}$$

$$y_2 = \begin{cases} 1 \; if \; \mathrm{y}_2^* > 0 \\ 0 \; if \; \mathrm{y}_2^* \leq 0 \end{cases}$$

- Marginal effects and predicted values can be estimated similarly to those for the binary probit models. Marginal effects for the joint probability, say P(y_1=1 and y_2=1) are also available.

Multinomial Probit and Logit Models
Conditional Logit Model
Mixed Logit Model

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Multinomial, Conditional, and Mixed Models Overview

- Multinomial outcome dependent variable (in wide and long form of data sets)
- Independent variables (alternative-invariant or alternative-variant)
- Multinomial logit model (coefficients, marginal effects, IIA) and multinomial probit model
- Conditional logit model (coefficients, marginal effects)
- Mixed logit model

Multinomial, Conditional and Mixed Models

Multinomial outcome examples

- The type of insurance contract that an individual selects.
- The product that an individual selects (say type of cereal).
- Occupational choice by an individual (business, academic, non-profit organization).
- The choice of fishing mode (beach, pier, private boat, charter boat).

Multinomial outcome dependent variable

- The dependent variable y is a categorical, unordered variable.
- An individual may select only one alternative.
- The choices/categories are called alternatives and are coded as j =1, 2, …, m.
- The numbers are only codes and their magnitude cannot be interpreted (use frequency for each category instead of means to summarize the dependent variable).
- The data are usually recorded in two formats: a wide format and a long format.
- When using the wide format, the data for each individual i is recorded on one row. The dependent variable is:

$$y = j$$

- When using the long format, the data for each individual *i* is recorded on *j* rows, where *j* is the number of alternatives. The dependent variable is:

$$y_j = \begin{cases} 1 \text{ if } y = j \\ 0 \text{ if } y \neq j \end{cases}$$

- Therefore, $y_j = 1$ if the alternative *j* is the observed outcome and the remaining $y_k = 0$. For each observation only one of y_1, y_2, …, y_m will be non-zero.

Example for multinomial data in wide form

Person ID (*i*)	Dependent variable (y)	Codes for y	w_i (income)	x_{i1} (price of alternative 1)	x_{i2} (price of alternative 2)
1	apple juice (alternative 1)	y=1	40,000	2.5	1.5
2	orange juice (alternative 2)	y=2	38,000	2.7	1.7
3	orange juice (alternative 2)	y=2	50,000	2.9	1.6

Example for multinomial data in long form

Person ID (*i*)	Dependent variable (y_j)	Codes for y_j	w_i (income)	x_{ij} (price)
1	apple juice (alternative 1)	$y_1 = 1$	40,000	2.5
1	orange juice (alternative 2)	$y_2 = 0$	40,000	1.5
2	apple juice (alternative 1)	$y_1 = 0$	38,000	2.7
2	orange juice (alternative 2)	$y_2 = 1$	38,000	1.7
3	apple juice (alternative 1)	$y_1 = 0$	50,000	2.9
3	orange juice (alternative 2)	$y_2 = 1$	50,000	1.6

- The multinomial density for one observation is defined as:

$$f(y) = p_1^{y_1} \times \dots \times p_m^{y_m} = \prod_{j=1}^{m} p_j^{y_j}$$

- The probability that individual i chooses the jth alternative is:

$$p_{ij} = \mathrm{pr}[y_i = j] = \mathrm{F}_j(\mathbf{x}_i, \beta)$$

- The functional form of F_j should be selected so that the probabilities lie between 0 and 1 and sum over j to one. Different functional forms of F_j lead to multinomial, conditional, mixed, and ordered logit and probit models.

Independent variables

- Two types of independent variables.
- *Alternative-invariant or case-specific regressors* –the regressors w_i vary over the individual *i* but do not vary over the alternative *j*.
    - Income, age, and education are different for each individual but they do not vary based on the type of a product that the individual selects.
    - Used in the multinomial logit model.
- *Alternative-variant or alternative-specific regressors* – the regressors x_{ij} vary over the individual *i* and the alternative *j*.
    - Prices for products vary for each product and individuals may also pay different prices.
    - Salaries for occupation may be different between occupations and also for each individual.
    - Used in the conditional and mixed logit models.

Multinomial logit model

- The multinomial logit model is used with alternative-invariant regressors.
- The probability that individual *i* will select alternative *j* is:

$$p_{ij} = p(y_i = j) = \frac{\exp\,(\mathbf{w}_i'\gamma_j)}{\sum_{\mathrm{k}=1}^{\mathrm{m}} \exp\,(\mathbf{w}_i'\gamma_k)}$$

- This model is a generalization of the binary logit model.
- The probabilities for choosing each alternative sum up to 1, $\sum_{\mathrm{j}=1}^{\mathrm{m}} p_{ij} = 1$
- One set of coefficients needs to be normalized to zero to estimate the models (usually $\gamma_1 = 0$), so there are (j-1) sets of coefficients estimated. The coefficients of other alternatives are interpreted in reference to the base outcome.
- Coefficient interpretation for alternative *j*: in comparison to the base alternative, an increase in the independent variable makes the selection of alternative *j* more or less likely.

Marginal effects

- The marginal effect of an increase of a regressor on the probability of selecting alternative j is:

$$\partial p_{ij}/\partial \mathbf{w}_i = p_{ij}(\gamma_j - \overline{\gamma_i})$$

- The marginal effects do not necessarily correspond in sign to the coefficients (unlike the binary logit or probit model).
- There are (j-1) sets of coefficients because one set is normalized to zero, but there are j sets of marginal effects.
- Depending on which alternative we select as a base category, the coefficients will be different (in reference to the base category) but the marginal effects will be the same regardless of the base category.
- The marginal effects of each variable on the different alternatives sum up to zero.
- Marginal effects interpretation: each unit increase in the independent variable increases/decreases the probability of selecting alternative j by the marginal effect expressed as a percent.

Independence from Irrelevant Alternatives (IIA) property

- The odds ratios in the multinomial logit models are independent of other alternatives. For choices *j* and *k*, the odds ratio only depends on the coefficients for choices *j* and *k*.
- Odds ratio: $p_{ij}/p_{ik} = \exp\left(\mathbf{w}_i'(\gamma_j - \gamma_k)\right)$
- This weakness of the multinomial model is known as the red bus-blue bus problem. If the choice is between a car and a blue bus, according to the model the introduction of a red bus will not change the probabilities.

Multinomial probit model

- The multinomial probit model is similar to multinomial logit model, just like the binary probit model is similar to the binary logit model.
- The difference is that it uses the standard normal cdf.
- The probability that observation *i* will select alternative *j* is:

$$p_{ij} = p(y_i = j) = \Phi\,(\mathbf{x}_{ij}'\beta)$$

- It takes longer for a probit model to obtain results.
- The coefficients are different by a scale factor from the logit model.
- The marginal effects will be similar.

Conditional logit model

- The conditional logit model is used with alternative-invariant and alternative-variant regressors.
- The probability that observation *i* will choose alternative j is:

$$p_{ij} = p(y_i = j) = \frac{\exp\left(\mathbf{x}_{ij}'\beta + \mathbf{w}_i'\gamma_j\right)}{\sum_{\mathrm{k=1}}^{\mathrm{m}} \exp\left(\mathbf{x}_{ik}'\beta + \mathbf{w}_i'\gamma_k\right)}$$

  where $\mathbf{x}_{ij}$ are alternative-specific regressors and $\mathbf{w}_i$ are case-specific regressors.
- The conditional logit model has (j-1) sets of coefficients (γ_j) (with one set being normalized to zero) for the case-specific regressors and only one set of coefficients (β) for the alternative-specific regressors.
- The probabilities for choosing each alternative sum up to 1.
- Coefficients for the alternative-invariant regressors γ_j (similar treatment as the multinomial logit model).
  - One set of coefficients for the alternative-invariant regressors is normalized to zero (say $\gamma_1 = 0$), this is the base outcome. The rest of coefficients are interpreted in relation to this base category.

    - There are (j-1) sets of coefficients (corresponding to the number of alternatives minus 1 for the base).
    - Coefficient interpretation for alternative j: in comparison to the base alternative, an increase in the independent variable makes the selection of alternative j more or less likely.
- Coefficients for the alternative-specific regressors (β).
    - No normalization is needed.
    - One set of coefficients across all alternatives.
    - Coefficient interpretation: an increase in the price of one alternative decreases the probability of choosing that alternative and increases the probability of choosing other alternatives.

Marginal effects

- The marginal effect of an increase of a regressor on the probability of selecting alternative *j* is:

$$\partial p_{ij}/\partial \mathbf{x}_{ik} = p_{ij}(\delta_{ijk} - p_{ik})\beta$$

  where $\delta_{ijk} = 1$ if j=k and 0 otherwise.

- There are *j* sets of marginal effects for both the alternative-specific and case-specific regressors.
- For each alternative-specific variable $\mathbf{x}_{ij}$, there are *j*x*j* sets of marginal effects.
- The marginal effects of each variable on the different alternatives sum up to zero.
- Marginal effects interpretation: each unit increase in the independent variable increases the probability of selecting the *k*th alternative and decreases the probability of the other alternatives, by the marginal effect expressed as a percent.

Mixed logit model

The mixed logit model (also called random parameters logit model) specifies the utility to the ith individual for the jth alternative to be:

$$U_{ij} = \mathbf{x}_{ij}'\beta_i + \mathbf{w}_i'\gamma_{ji} + e_{ij} = \mathbf{x}_{ij}'\beta + \mathbf{w}_i'\gamma_j + \mathbf{x}_{ij}'\upsilon_i + \mathbf{w}_i'\delta_{ji} + e_{ij}$$

where e_{ij} are iid extreme value (similar to the errors in the conditional logit model).

- The mixed logit model allows for the parameters β_i to be random. A common assumption is that $\beta_i = \beta + \upsilon_i$ where $\upsilon_i \sim \mathrm{N}[0, \Sigma_\beta]$ and $\gamma_{ji} = \gamma_j + \delta_{ji}$ where $\delta_{ji} \sim \mathrm{N}[0, \Sigma_{\gamma i}]$.
- The introduction of the random parameters has the attractive property of inducing correlation across alternatives. The combined error $\mathbf{x}_{ij}'\upsilon_i + \mathbf{w}_i'\delta_{ji} + e_{ij}$ is now correlated across alternatives, say $\mathrm{Cov}\left[\upsilon_{ij}, \upsilon_{ij}\right] = \mathbf{x}_{ij}'\Sigma_\beta\mathbf{x}_{ik}$.
- The probability that individual i selects alternative j represents a mixed logit model:

$$p_{ij} = p(y_i = j) = \frac{\exp\left(\mathbf{x}_{ij}'\beta + \mathbf{w}_i'\gamma_j + \mathbf{x}_{ij}'\upsilon_i + \mathbf{w}_i'\delta_{ji}\right)}{\sum_{\mathrm{k=1}}^{\mathrm{m}} \exp\left(\mathbf{x}_{ik}'\beta + \mathbf{w}_i'\gamma_k + \mathbf{x}_{ik}'\upsilon_i + \mathbf{w}_i'\delta_{ki}\right)}$$

- The mixed logit model relaxes the IIA assumption by allowing parameters in the conditional logit model to be normally (or log-normally) distributed.
- When estimating the mixed logit model, the researcher needs to specify which parameters will be estimated as random. If a parameter is random, this implies that effect of a particular regressor on the chosen alternative varies across the individuals.
- The mixed logit model produce random parameters coefficients for both the regressor (x_i) and the standard deviation of the regressor (sd(x_i)).
- Coefficient interpretation for the regressors (x_i): when the independent variable increases, the consumers are more or less likely to choose this alternative.
- Coefficient interpretation on the standard deviation of a regressor (sd(x_i)): there is a heterogeneity across individuals with respect to the effect of the independent variable on the alternative chosen.

Ordered Probit and Logit Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Ordered Probit and Logit Models Overview

- Ordered outcome dependent variable
- Ordered probit and logit models (coefficients and marginal effects)

Ordered Probit and Logit Models

Ordered outcome examples

- Rating systems (poor, fair, good, excellent)
- Opinion surveys (strongly agree, agree, neutral, disagree, strongly disagree).
- Employment (unemployed, part time, full time)
- Ranking (senior, junior, sophomore, freshman)
- Grades (A, B, C, D, E)
- Bond ratings (AAA, AA, A, B, etc.)

Ordered outcome dependent variable

- The categories for the dependent variables are rankings so the numbers do not make sense (even if they are coded as 0, 1, 2, 3, 4, the difference between the first and second outcome may not be the same as between the second and third).

Ordered logit and probit models

- An index model for a single latent variable y* (which is unobservable, we only know when it crosses thresholds).

$$y_i^* = \mathbf{x}_i'\beta + u_i$$

$$y_i = j \quad \text{if} \quad \alpha_{j-1} < y_i^* \leq \alpha_j$$

- The probability that observation i will select alternative j is:

$$p_{ij} = p(y_i = j) = p\left(\alpha_{j-1} < y_i^* \leq \alpha_j\right) = F\left(\alpha_j - \mathbf{x}_i'\beta\right) - F\left(\alpha_{j-1} - \mathbf{x}_i'\beta\right)$$

- For the ordered logit, *F* is the logistic cdf $F(z) = e^z/(1 + e^z)$.
- For ordered probit, F is the standard normal cdf.

- The ordered logit/probit model with *j* alternatives will have one set of coefficients with (*j*-1) intercepts. You can recognize an ordered choice model by the multiple intercepts.
- The ordered logit/probit model with *j* alternatives will have *j* sets of marginal effects.

- Interpretation of coefficients: the sign of parameters shows whether the latent variable y* increases with the regressor. The magnitude of the coefficients will be different by a scale factor between the probit and logit models.

Marginal effects for the ordered logit/probit models

- The marginal effect of an increase in a regressor x_r on the probability of selecting alternative j is:

$$\partial p_{ij}/\partial \mathbf{x}_{ri} = \{F'(\alpha_{j-1} - \mathbf{x}_i'\beta) - F'(\alpha_j - \mathbf{x}_i'\beta)\}\beta_r$$

- The marginal effects of each variable on the different alternatives sum up to zero.
- Marginal effects interpretation: each unit increase in the independent variable increases/decreases the probability of selecting alternative j by the marginal effect expressed as a percent.

Limited Dependent Variable Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Limited Dependent Variable Models Overview

- Limited dependent variable examples
- Censoring and truncation
- Tobit model
- Truncated regression
- Heckman model

Limited Dependent Variable Models

Limited dependent variable model explanations and examples

- A limited dependent variable means that there is a limit or boundary on the dependent variable and some of the observations “hit” this limit.
- Consumer economics: the quantity of product consumed is zero for some consumers and positive amounts for the rest.
- Labor economics: the labor supply (hours worked) by women with some women choosing not to work (zero hours) and others choosing to work a positive number of hours.
- Capacity issues: the demand for tickets for a game or conference is censored at the hall capacity (the max tickets that can be sold).
- Top coding: annual income data can be top-coded at 100,000 (recording 100,000 for all observations with incomes above 100,000).

Censoring and truncation

- Censoring is when the limit observations are in the sample (only the value of the dependent variable is censored) and truncation is when the observations are not in the sample.
    - Censored sample: include consumers who consume zero quantities of a product.
    - Truncated sample: only include consumers who choose positive quantities of a product.
- The censored sample is representative of the population (only the mean for the dependent variable is not) because all observations are included. The truncated sample is not representative of the population because some observations are not included.
- Truncation has greater loss of information than censoring (missing observations rather than values for the dependent variable).
    - Censored sample: observe people that do not work but their work hours are recorded as zero.
    - Truncated sample: do not observe anything about people who do not work.
- A truncated sample will have fewer observations and higher mean (with censoring from below) than a censored sample.
- Because of censoring, the dependent variable y is the incompletely observed value of the latent dependent variable y^*.
    - Income of $y^*=120,000$ will be censored as $y=100,000$ with top coding of 100,000.

Censoring from below

- The actual value for the dependent variable *y* is observed if the latent variable *y** is above the limit and the limit is observed for the censored observations.
  - We observe the actual hours worked for people who work and zero for people who do not work.

$$y = \begin{cases} y^* \; if \; y^* > L \\ L \; if \; y^* \leq L \end{cases}$$

where *L* is the lower limit.

Censoring from above

- The actual value for the dependent variable *y* is observed if the latent variable *y** is below the limit and the limit is observed for the censored observations.
  - If people make below $100,000, we observe their actual income and if they make above $100,000, we record their income as 100,000 (censored values).

$$y = \begin{cases} y^* \; if \; y^* < U \\ U \; if \; y^* \geq U \end{cases}$$

where *U* is the upper limit.

Truncation from below

- The dependent variable is continuous and no zeros are allowed (truncated at zero).
- The actual values for the dependent variable are observed if they are greater than the lower limit.
  - We observe the working hours for a sample of working women or the quantity of product consumed for a sample of people with positive consumption.

$$y = y^* \; if \; y^* > L$$

Truncation from above

- The actual values for the dependent variable are observed if they are lower than the upper limit.
  - We observe the actual income in a sample of low-income people.

$$y = y^* \; if \; y^* < U$$

Tobit model

- The Tobit model is the censored normal regression model:

$$y^* = \mathbf{x}'\beta + e$$

- The most common Tobit model is when the dependent variable is censored from below at zero (lots of zeros and positive values in the sample).

$$y = \begin{cases} y^* if\ y^* > 0 \\ 0\ if\ y^* \leq 0 \end{cases}$$

- The dependent variable can also be expressed as:

$$y = \max{(y^*, 0)}$$

- The upper limit, lower limit, or two-limit values will need to be specified to estimate the models.
- If no observations are censored, the Tobit model is the same as an OLS regression.

Tobit model = Probit + truncated regression

- The Tobit model is a combination of two models (decisions):

1. Probit model for the discrete decision of whether or not *y* is zero or positive:

$$Prob(y > 0) = \Phi(\mathbf{x}'\beta)$$

2. Truncated regression model for the continuous decision (for the quantity of *y*|*y*>0).

$$E(y|y > 0) = \mathbf{x}'\beta + \sigma\lambda\left(\frac{\mathbf{x}'\beta}{\sigma}\right)$$

- In the Tobit model, the coefficients on the probit model and on the truncated regression are restricted to be the same.
- The Tobit model assumes normality like the probit model.

Marginal effects

Marginal effects for the latent variable

- The marginal effects for the latent variable are the coefficients.
  - Marginal effect on the desired hours of work.

$$dE(y^*)/dx = \beta$$

Marginal effects for the censored sample (Tobit model)

- For the censored sample (with zeros and positive amounts)
  - Marginal effect on the actual hours of work for workers and non-workers.

$$dE(y)/dx = \beta\Phi\left(\mathbf{x}'\beta/\sigma\right)$$

- The marginal effects in the Tobit model are the coefficients multiplied by a positive scale factor.

Marginal effects for the truncated sample

- For the truncated sample (with positive amounts)
  - Marginal effect on the actual hours of work for workers.

$$dE(y|y > 0)/dx = \left[1 - \frac{\mathbf{x}'\beta}{\sigma}\lambda\left(\frac{\mathbf{x}'\beta}{\sigma}\right) - \lambda\left(\frac{\mathbf{x}'\beta}{\sigma}\right)^2\right]\beta$$

- The marginal effects show the highest impact of the independent variables for the latent variable, less impact for the censored sample, and even less impact for the truncated sample.
- We use the different marginal effects depending on what variable is of interest in the study.
  - In a top-coded study of income, the effect on the latent variable will be of interest (we do not care about the censoring at $100,000).

Cragg’s model (two-part, hurdle model)

- Cragg’s model relaxes the restrictive assumption of the Tobit model that the discrete decision and the continuous decision are the same.
  - Example: older buildings are more likely to catch fire and claim insurance payments, but the payments will be smaller because older buildings are cheaper. We need a positive effect of the building age on the probability of a payment, and negative effect of the building age on the amount of payment.
  - Cragg’s model can show this effect, but the Tobit model cannot.
- Cragg’s model is also called a hurdle model.

Cragg’s model two-step estimation procedure

1. Probit model for the discrete decision:

$$Prob(y^* > 0) = \Phi(\mathbf{x}'\gamma)$$

2. Truncated regression model for the continuous decision (uncensored observations):

$$E(y|y^* > 0) = \mathbf{x}'\beta + \sigma\lambda\left(\frac{\mathbf{x}'\beta}{\sigma}\right)$$

- Note that in the Cragg's two-step model, the coefficients in the two steps can be different (γ) and (β). We can also have different sets of variables x and z in the first and second step of the model.

Test for the Tobit one-step model versus the Cragg's two-step model

- We have the Tobit model if the restriction $\gamma = \beta/\sigma$ holds (same coefficients for the discrete and continuous decisions).
- We have the Cragg's model if the restriction $\gamma = \beta/\sigma$ is rejected (different coefficients for the discrete and continuous decisions).
- To test this restriction, estimate separately Tobit, probit, and truncated regression models and get their log likelihoods. Then compute the following likelihood ratio statistic:

$$\lambda = 2 * (LL_{Probit} + LL_{Truncreg} - LL_{Tobit})$$

- The test statistic has a chi-square distribution with degrees of freedom equal to the number of independent variables (including an intercept). The Tobit model is rejected in favor of Cragg's model if λ exceeds the appropriate chi-square critical value.

Heckman model

- The Heckman model is a sample selection model.
- Sample selection usually occurs when people select themselves into a group.
  - We want to study the factors affecting income for working women. We have the selection decision, whether women choose to work or not. We also have the income only for women that work.
  - Different factors may affect the two decisions. Whether women work or not may be influenced by whether or not they have kids, but their income should not be influenced by the presence of kids.
- Sample selection (incidental truncation) is different from truncation.
  - We want to study income. Truncation is if the sample is based on high income. Incidental truncation is if the sample is based on whether or not people have executive jobs (high correlation but not exactly the same).
- The dependent variable is not observed if the observation is not in the sample.
  - We do not know the income for people who are not in the high income sample (but the income is not zero).
- Sample selection assumes that the discrete decision z and the continuous decision y have a bivariatec distribution with correlation ρ.

Heckman model two-step estimation procedure

1. Probit model for the selection mechanism

$$Prob(z = 1) = \Phi(\mathbf{w}'\gamma)$$

- Compute the inverse Mills ratio

$$\hat{\lambda}(\mathbf{w}'\gamma) = \frac{\phi(\mathbf{w}'\hat{\gamma})}{\Phi(\mathbf{w}'\hat{\gamma})}$$

2. Regression model for the selected sample

$$E(y|z = 1) = \mathbf{x}'\beta + \rho\sigma\hat{\lambda}(\mathbf{w}'\gamma)$$

- The Heckman model may or may not have the same regressors for the selection equation and regression.
- The Heckman model will report estimates of λ, ρ, σ.

Count Data Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Count Data Models Overview

- Count data examples
- Count data dependent variable
- Poisson model (properties of the distribution, coefficients, and marginal effects)
- Negative binomial model (test for overdispersion and incidence rate ratios)
- Hurdle or two-part models
- Zero-inflated models

Count Data Models

Count data examples

- Consumer demand: the number of products that a consumer buys on Amazon
- Recreational data: the number of trips taken per year
- Family economics: the number of children a couple has
- Health demand: the number of doctors visits

Count data dependent variable

- The dependent variable is counts (a non-negative integer): y = 0, 1, 2, 3, 4, …
- The sample is concentrated on a few small discrete values.
- We study the factors affecting the average number of the dependent variable.

Poisson model

Poisson model

- The Poisson model predicts the number of occurrences of an event.
- The Poisson model states that the probability that the dependent variable *Y* will be equal to a certain number *y* is:

$$p(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- For the Poisson model, μ is the intensity or rate parameter.

$$\mu = \exp\left(\mathbf{x}_i^{'}\beta\right)$$

- Interpretation of the coefficients: one unit increase in x will increase/decrease the average number of the dependent variable by the coefficient expressed as a percentage.

Properties of the Poisson distribution

- *Equidispersion property* of the Poisson distribution: the equality of mean and variance.

$$E(y|x) = var(y|x) = \mu$$

  This is a restrictive property and often fails to hold in practice, i.e., there is “overdispersion” in the data. In this case, use the negative binomial model.

- *Excess zeros problem* of the Poisson distribution: there are usually more zeros in the data than a Poisson model predicts. In this case, use the zero-inflated Poisson model.

Marginal effects for the Poisson model

- The marginal effect of a variable on the average number of events is:

$$\partial E(y|x)/\partial x_j = \beta_j \exp\,(\mathbf{x}_i^{'}\beta)$$

- Interpretation of the marginal effects: one unit increase in x will increase/decrease the average number of the dependent variable by the marginal effect.

Negative binomial model

- The negative binomial model is used with count data instead of the Poisson model if there is overdispersion in the data.
- Unlike the Poisson model, the negative binomial model has a less restrictive property that the variance is not equal to the mean (μ).

$$var(y|x) = \mu + \alpha\mu^2$$

- Another functional form is $var(y|x) = \mu + \alpha\mu$, but this form is less used.
- The negative binomial model also estimates the overdispersion parameter α.

Test for overdispersion

- Estimate the negative binomial model which includes the overdispersion parameter α and test if α is significantly different than zero.
- H_0: $\alpha = 0$ or H_a: $\alpha \neq 0$
- We have three cases:
  - When $\alpha = 0$, the Poisson model.
  - When $\alpha > 0$, overdispersion (frequently holds with real data).
  - When $\alpha < 0$, underdispersion (not very common).

Incidence rate ratios (irr)

- For the Poisson and negative binomial models, in addition to reporting the coefficients and marginal effects, we can also report the incidence rate ratios.
- The incidence rate ratios report exp(b) rather than b.
- Interpretation of the incidence rate ratios: irr=2 means that for each unit increase in x, the expected number of y will double.

Hurdle or two-part models

- The two-part model relaxes the assumption that the zeros (whether or not there are events) and positives (how many events) come from the same data generating processes.
- Example: different factors may affect whether or not you practice a particular sport and how many times you practice your sport in a month.
- We can estimate two-part models similar to the truncated regression models.
- If the process generating the zeros is $f_1(.)$ and the process generating the positive responses is $f_2(.)$ then the two-part hurdle model is defined by the following probabilities:

$$g(y) = \begin{cases} f_1(0) \ \ if\ y = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y) \ \ if\ y \geq 1 \end{cases}$$

- If the two processes are the same, this is the standard count data model.
- The model for the zero versus positive responses is a binary model with the specified distribution, but we usually estimate it with the probit/logit model.

Zero-inflated models

- The zero-inflated model is used with count data when there is an excess zeros problem.
- The zero-inflated model lets the zeros occur in two different ways: as a realization of the binary process (z=0) and as a realization of the count process when the binary variable z=1.
- Example: you either like hiking or you do not. If you like hiking, the number of hiking trips you can take is 0, 1, 2, 3, etc. So you may like hiking, but may not take a trip this year. We are able to generate more zeros in the data.
- If the process generating the zeros is $f_1(.)$ and the process generating the positive responses is $f_2(.)$ then the zero-inflated model is:

$$g(y) = \begin{cases} f_1(0) + (1 - f_1(0))f_2(0) \;\; if\ y = 0 \\ (1 - f_1(0))f_2(y) \;\; if\ y \geq 1 \end{cases}$$

- The zero-inflated model is less frequently used than the hurdle model.
- The zero-inflated models can handle the excess zeros problem.

Survival Analysis

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Survival Analysis Overview

- Survival analysis examples
- Survival analysis set up and features
- Extensions of basic survival analysis
- Survival, hazard, and cumulative hazard functions
- Nonparametric analysis (Kaplan-Meier survival function)
- Parametric models (Exponential, Weibull, Gompertz, and Log-logistic)
- Semi-parametric models (Cox proportional hazard model)

Survival Analysis

Survival analysis is also called duration analysis, transition analysis, failure time analysis, and time-to-event analysis.

Survival analysis examples

- Finance: Loan performance (borrowers obtain loans and then they either default or continue to repay their loans)
- Economics: Firm survival and exit
- Economics: Time to retirement, finding a new job, etc.
- Economics: Adoption of new technology (firms either adopt the new technology or still haven’t adopted it)

Survival analysis set up

- Subjects are tracked until an event happens (failure) or we lose them from the sample (censored observations).
- We are interested in how long they stay in the sample (survival).
- We are also interested in their risk of failure (hazard rates).

Survival analysis features

- The dependent variable is duration (time to event or time to being censored) so it is a combination of time and event/censoring.
  - time variable = length of time until the event happened or as long as they are in the study
  - the event variable = 1 if the event happened or 0 if the event has not yet happened
  - Instead of an event variable, a censor variable can be defined. The censored variable =1 if the event has not happened yet, and 0 if the event has happened.

Time	Event/ Failure	Censored	Explanation
15	0	1	Event hasn't happened yet (censored)
22	1	0	Event happened (not censored)
78	0	1	Event hasn't happened yet (censored)
34	1	0	Event happened (not censored)

- The hazard rate is the probability that the event will happen at time *t* given that the individual is at risk at time *t*.
- Hazard rates usually change over time.
  - The probability of defaulting on a loan may be low in the beginning but increases over the time of the loan.

Extensions of the basic survival analysis

- Multiple occurrences of events (multiple observations per individual)
    - borrower may have repeated restructuring of the loan
    - firm may adopt technology in some years but not others
- More than one type of event (include codes for events, e.g. 1, 2, 3, 4)
    - borrower may default (one type of event) or repay the loan earlier (a second type of event)
    - firms may adopt different types of technologies
- Two groups of participants
    - the effect of two types of educational programs on technology adoption rates
- Time-varying covariates
    - borrower's income may have changed during the study which caused the default.
- Discrete instead of continuous transition times
    - exits are measured in intervals (such as every month)
- There may different starting times – we need to measure time from the beginning time to the event.

Survival, hazard, and cumulative hazard functions

- The dependent variable duration is assumed to have a continuous probability distribution $f(t)$.
- The probability that the duration time will be *less than t* is:

$$F(t) = Prob(T \leq t) = \int_0^t f(s)ds$$

- *Survival function* is the probability that the duration will be *at least t*:

$$S(t) = 1 - F(t) = Prob(T \geq t)$$

- *Hazard rate* is the probability that the duration will end after time *t*, given that it has lasted until time *t*:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

- The hazard rate is the probability that an individual will experience the event at time *t* while that individual is at risk for experiencing the event.

Nonparametric models

- Nonparametric estimation is useful for descriptive purposes and to see the shape of the hazard or survival function before a parametric model with regressors is introduced.

Time t_j	Number at risk n_j	Number of events d_j	Number of censored observations	Hazard function $\lambda = d_j/n_j$	Cumulative hazard function $\Lambda(t_j)$	Survival function $S(t_j)$
3	100	10	3	10/100=0.1	0.1	1-0.1=0.9
4	100-10-3=87	3	2	3/87=0.034	0.1+0.034 =0.134	0.9*(1-0.034) =0.87
5	87-3-2=82	6	1	6/82=0.073	0.134+0.073 =0.207	0.87*(1-0.073)=0.81

- Think about the shapes of the hazard function and survival function plotted over time.

Survival analysis nonparametric procedure

- Sort the observations based on duration from smallest to largest $t_1 \leq t_2 \leq \cdots \leq t_n$
- For each duration, determine the number of observations at risk n_j (those still in the sample), the number of events d_j and the number of censored observations m_j.
- Calculate the hazard function as the number of events as a proportion of the number of observations at risk

$$\lambda(t_j) = \frac{d_j}{n_j}$$

- *Nelson-Aalen estimator of the cumulative hazard function* – calculated by summing up hazard functions over time:

$$\Lambda(t_j) = \sum \frac{d_j}{n_j}$$

- *The Kaplan-Meier estimator of the survival function* – take the ratios of those without events over those at risk and multiply that over time.

$$S(t_j) = \prod \frac{n_j - d_j}{n_j}$$

A few facts about the Kaplan-Meier survival function

- It is a decreasing step function with a jump at each discrete event time.
- Without censoring, the Kaplan-Meier estimator is just the empirical distribution of the data.

Parametric and semiparametric models

- Unlike the nonparametric estimation, the parametric models also allow the inclusion of independent variables.

Parametric models

- Parametric models can assume different parametric forms for the hazard function.

Parametric model	Hazard function λ	Survival function S
Exponential	γ	$\exp(-\gamma t)$
Weibull	$\gamma \alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$
Gompertz	$\gamma \exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$
Log-logistic	$\alpha\gamma^{\alpha} t^{\alpha-1}/(1+(\gamma t)^{\alpha})$	$1/(1+(\gamma t)^{\alpha})$

- The exponential model has a constant hazard rate over time.

Cox proportional hazard model

- The hazard rate in the Cox proportional hazard model is defined as:

$$\lambda(t|\mathbf{x},\beta) = \lambda_0(t)\exp\left(\mathbf{x}'\beta\right)$$

Estimation of the parametric models

- For the parametric and semiparametric models, report both the coefficients and hazard ratios.
- Interpretation of coefficients: a positive coefficient means that as the independent variable increases the time-to-event *decreases*, (lower duration or more likely for the event to happen).
- Interpretation of hazard rates: a hazard ratio of 2 (0.5) means that for a one unit increase in the x variable, the hazard rate (probability of event happening) increases by 100% (decreases by 50%). A hazard rate of greater than 1 means that it is more likely for the event to happen.

Coefficient	Hazard rate	Conclusion
Positive	>1	Lower duration, higher hazard rates (more likely for the event to happen).
Negative	(0,1)	Higher duration, lower hazard rates (less likely for the event to happen).

Spatial Econometrics

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Spatial Econometrics Overview

- Examples of spatial econometrics models and data
- Spatial weight matrix based on contiguity and distance
- Spatial regression models: spatial lag model and spatial error model
- Spatial dependence test (Moran’s I)

Spatial Econometrics

Spatial econometrics examples

- Real estate economics: House prices depend on the number of bedrooms, bathrooms, etc. House prices also depend on location; prices of houses in the same neighborhood are similar.
- Farmland values: Farmland prices depend on land rent, government payments, etc. but farmland prices are similar if counties are spatially close.
- Precision agriculture: Different rates of nitrogen are applied in a corn field. Corn yields will be different because of the different nitrogen applications but they will be similar if the fields/plots are spatially close.

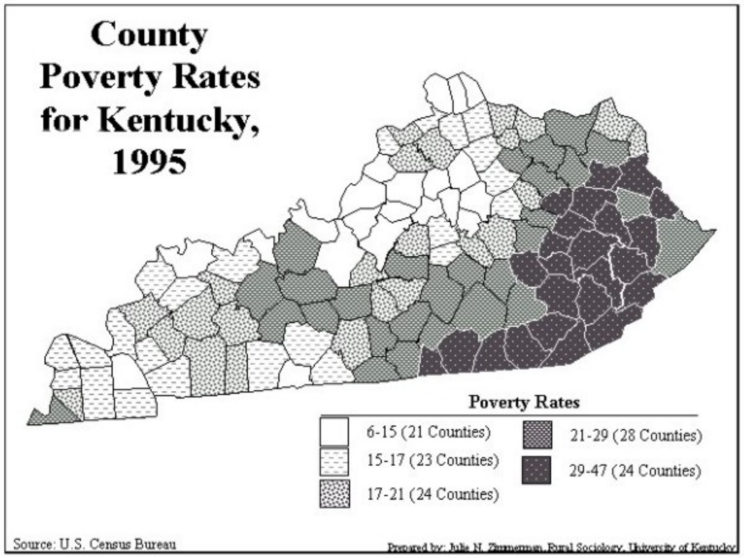
Spatial econometrics

- Spatial econometrics accounts for the presence of spatial effects in regression analysis.
- Spatial econometrics is used in regional science, urban and real estate economics and economic geography.

Spatial econometrics problems

- Spatial econometrics is a special type of econometrics.
- Spatial data needs to be geo-coded for location (coordinates, borders, distance).
- The spatial matrix defines neighbors (observations that are spatially close) and their effects.
- We need to account for the spatial dependence in the regression model.

Spatial data with spatial dependence (spatial autocorrelation)


County Poverty Rates for Kentucky, 1995
Poverty Rates
6-15 (21 Counties)
15-17 (23 Counties)
17-21 (24 Counties)
21-29 (28 Counties)
29-47 (24 Counties)
Source: U.S. Census Bureau
Prepared by: Julie N. Zimmerman, Rural Sociology, University of Kentucky

Spatial weight matrix

- The spatial weight matrix provides the structure of the spatial relationship among observations.
- The spatial weight matrix provides information about which observations are considered neighbors and also how their values are related to each other.
- The spatial weight matrix is defined as W with elements w_{ij} indicating whether observations i and j are spatially close.
- Spatial weight matrices need to be "row-standardized" which means the weights need to sum up to one on each row.
- There are two types of spatial weight matrices based on contiguity and on distance.

Spatial weight matrix based on contiguity

- We need to know if observations are contiguous (share a border and/or a vertex).
- Use GIS ArcView to specify the spatial weights (most software can't recognize contiguity).
- The spatial weight matrix W has elements defined as:

$$w_{ij} = \begin{cases} 1 \; if \;\; i \; is \; contiguous \; to \; j \\ 0 \; otherwise \end{cases}$$

- If units i and j are neighbors, the spatial weight is 1, otherwise 0.

- Example: spatial weight matrix based on *contiguity* where units 2 and 3 are neighbors and units 3 and 4 are neighbors (but units 2 and 4 are not neighbors).

$$W = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

- We can row-standardize the spatial weight matrix by dividing each entry by the total for that row.

$$W = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

- Now each row represents the unit's value as a weighted average of the values of its neighbors: y_3=0.5*y_2+0.5*y_4. In general, $y_i = \sum_j w_{ij} y_j$.
- The row standardization is needed because in a weighted average formula, the weights need to sum up to 1.
- We can predict the values for each unit based on the values of its neighbors.

Spatial weight matrix based on distance

Coordinates

- We need to know the location of the observations (X coordinates and Y coordinates (longitude and latitude)) to calculate the distance between observations.
- With county data, we use distance between the centroids (center points) of counties.
- Software can typically calculate the spatial weight matrix if the x and y coordinates are provided for each observation.

Distance

- Let d_{ij} be the distance between observations *i* and *j*.
- Assume that there are no spatial effects beyond a certain distance band *D*.

Spatial weight matrix

- A spatial weight matrix can be constructed based on distance where units within a specified radius have a spatial weight of 1 (they are neighbors), otherwise 0.

$$w_{ij} = \begin{cases} 1 \; if\ the\ distance\ between\ i\ and\ j < D \\ 0 \; otherwise \end{cases}$$

  - The spatial weight matrix will look similar to the one in the previous section – only it will be based on distance, not contiguity.

- Alternative specification: a spatial weight matrix can also be constructed based on distance where units with distance d_{ij} receive a weight that is inversely proportional to the distance between the units and 0 if they are beyond a certain distance band D.

$$w_{ij} = \begin{cases} 1/d_{ij} \; if\ the\ distance\ between\ i\ and\ j < D \\ 0 \; otherwise \end{cases}$$

- Example: spatial weight matrix based on *distance* where units 1 and 2 are neighbors with distance=5 and units 2 and 3 are neighbors with distance=2.

$$W = \begin{bmatrix} 0 & .2 & 0 & 0 \\ .2 & 0 & .5 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- The spatial weight matrix is row-standardized by dividing each entry by the total for that row. Now the value for each unit is a weighted average of its neighbors' values.

$$W = \begin{bmatrix} 0 & 1 & 0 & 0 \\ .3 & 0 & .7 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- Create the spatial map for the unit locations.

- The choice of a spatial weight matrix based on contiguity or distance is up to the researcher.
    - Both methods (contiguity and distance) are used in the literature.
    - Example: for farmland value, neighbors that are based on distance may be more appropriate, but for residential housing values, neighbors based on contiguity (or neighborhood block) may be more appropriate.

Spatial weight matrix properties

- The diagonal elements of the spatial matrix are set equal to zero and the non-diagonal elements are non-zero for observations that are spatially close to one another and zero for those that are far away.
- As the distance band goes to zero, the spatial regression results approximate those of OLS (no spatial effects).
- Spatial weight matrices are row standardized, meaning that the row elements sum up to 1. Each unit's value is a weighted average of its neighbors.
- The dimensions of a spatial matrix NxN are based on the sample size N. This is a major issue with large data sets.
- The researcher picks the distance band. If you set your distance band to the maximum distance, then all units will have at least one neighbor.

Spatial regression

- Spatial regression is a regression that accounts for the spatial dependence of the data.
- Spatial dependence is added to a regression in two ways: spatial lag and spatial error.

Spatial lag regression

- The spatial lag model is appropriate when the focus is on the spatial interactions of the dependent variable. In this case we know the structure of the spatial relationship.
    - Example: the price of a house will depend on the prices of neighboring houses.
- Here the dependent variable y has the spatial structure.
- The spatial lag model is a spatial autoregressive model that includes a spatially lagged dependent variable.
- The spatial lag of the variable y is Wy. The dependent variable is a weighted average of its neighbors' values.

- The spatial lag regression is defined as:

$$y = \rho W y + x\beta + e$$

- The spatial lag model reduced form equation is:

$$(I - \rho W)y = x\beta + e$$

- The independent variables are explaining the variation in the dependent variable that is not explained by the neighbors' values.
- The spatial dependence parameter ρ is also estimated.

Spatial error regression

- The spatial error model is appropriate when we are interested in correcting for spatial autocorrelation due to the use of spatial data (irrespective of whether the model of interest is spatial or not). In this case we do not know the structure of the spatial relationship.
- We include spatially correlated errors due to unobservable features or omitted variables associated with location.
  - Example: farmers' technology adoption decisions may be influenced by their neighbors.
- Here the error term e has the spatial structure.
- The regression model is:

$$y = x\beta + e$$

- The errors are spatially correlated:

$$e = \lambda We + u$$

or

$$(I - \lambda W)e = u$$

- The spatial error regression reduced form equation is:

$$(I - \lambda W)y = (I - \lambda W)x\beta + u$$

- The multipliers in front of the dependent and independent variables are the variation that cannot be explained by the neighbors' values.
- The spatial dependence parameter λ is also estimated.

Spatial dependence test (Moran's *I*)

- Moran's *I* test statistic is used to test if the data have spatial dependence.

$$I = (N/S_0)(e'We/e'e)$$

  - Here S_0 is a standardization factor that corresponds to the sum of weights for the non-zero cross-products: $S_0 = \sum_i \sum_j w_{ij}$
  - For row-standardized weights $S_0 = N$, so $I = e'We/e'e$.

Quantile Regression

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Quantile Regression Overview

- Quantile regression examples
- Quantile regression model
- Quantile regression coefficients and marginal effects
- Advantages of quantile regression

Quantile Regression

Quantile regression examples

- Consumer economics: effects of household income on food expenditures for low- and high-expenditure households.
- Education: factors affecting student scores along their score distribution.

Quantile regression model explanations

- We use quantiles to describe the distribution of the dependent variable.
- Quantiles and percentiles are synonymous – the 0.99 quantile is the 99th percentile.
- The best-known quantile is the median. The median is the 0.50 quantile.
- The standard Ordinary Least Squares (OLS) models the relationship between one or more independent variables *x* and the conditional mean of a dependent variable *y*.
- A quantile regression models the relationship between *x* and the conditional quantiles of *y* rather than just the conditional mean of *y*.
- A quantile regression gives a more comprehensive picture of the effect of the independent variables on the dependent variable.
- The dependent variable is continuous with no zeros or too many repeated values.

Quantile regression model

- The quantile regression is described by the following equation:

$$y_i = x_i^{'}\beta_q + e_i$$

where β_q is the vector of unknown parameters associated with the q^th quantile.

- The OLS minimizes $\sum_i e_i^2$, the sum of squares of the model prediction error e_i
- The median regression, also called least absolute-deviation regression minimizes $\sum_i |e_i|$
- The quantile regression minimizes $\sum_i q|e_i| + \sum_i (1-q)|e_i|$, a sum that gives the asymmetric penalties $q|e_i|$ for underprediction and $(1-q)|e_i|$ for overprediction

The qth quantile regression estimator $\widehat{\beta_q}$ minimizes over β_q the objective function

$$Q(\beta_q) = \sum_{i:y_i \geq x_i^{'}\beta}^{N} q|y_i - x_i^{'}\beta_q| + \sum_{i:y_i < x_i^{'}\beta}^{N} (1-q)|y_i - x_i^{'}\beta_q|$$

where $0 < q < 1$.

- In contrast to OLS and maximum likelihood, the quantile regression computational implementation uses linear programming methods.
- We have β_q instead of β to make clear that different choices of q estimate different values of β.

Quantile regression coefficients and marginal effects

- The standard conditional quantile is specified to be linear:

$$Q_q(y_i|x_i) = x_i^{'}\beta_q$$

- For the *j*th regressor, the marginal effect is the coefficient for the qth quantile.

$$\frac{\partial Q_q(y|x)}{\partial x_j} = \beta_{qj}$$

- A quantile regression parameter β_{qj} estimates the change in a specified quantile q of the dependent variable *y* produced by a one unit change in the independent variable x_j.
- The marginal effects are for infinitesimal changes in the regressor, assuming that the dependent variable remains in the same quantile.

- Unlike interpretations of OLS regression, the interpretations of quantile regression results need to specify which quantile of the dependent variable they refer to.
- Two types of significance are important for quantile regression coefficients.
    - Quantile coefficients can be significantly different from zero.
    - Quantile coefficients can be significantly different than the OLS coefficients, showing different effects along the distribution of the dependent variable.

Advantages of the quantile regression

- Flexibility for modeling data with heterogeneous conditional distributions.
- Median regression is more robust to outliers than the OLS regression.
- Richer characterization and description of the data: can show different effects of the independent variables on the dependent variable depending across the spectrum of the dependent variable.

Propensity Score Matching

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Propensity Score Matching Overview

- Treatment evaluation examples and definitions
- Propensity score methodology
    - Treated and control groups
    - Probit/logit models to estimate propensity
    - Matching methods for treated and controlled observations
    - Treatment effects estimation
- Assumptions in propensity score matching
- Difference-in-differences models

Propensity Score Matching

Treatment evaluation definition

- Treatment evaluation is the estimation of the average effects of a program or treatment on the outcome of interest.
- Comparison of outcomes between treated and control observations.

Treatment evaluation examples

- Effects of training programs on job performance
- Government programs targeted to help schools and their effect on student performance

Two types of studies

- controlled experiments (assignment into treated and control groups is random)
- observational studies (assignment into treated and control groups is not random)

Propensity score matching methodology

- Assign the observations into two groups: the treated group that received the treatment and the control group that did not.
    - Treatment D is a binary variable that determines if the observation has the treatment or not
    - *D*=1 for treated observations and *D*=0 for control observations

- Estimate a probit/logit model for the propensity of observations to be assigned into the treated group. Use *x* variables that may affect the likelihood of being assigned into the treated group.
    - The propensity score model is a probit/logit model with *D* as the dependent variable and *x* as independent variables.

$$p(x) = prob(D = 1|x) = E(D|x)$$

    - The propensity score is the conditional (predicted) probability of receiving treatment given pre-treatment characteristics *x*.

- Match observations from treated and control groups based on their propensity scores
    - Several matching methods are available: kernel, nearest neighbor, radius, stratification

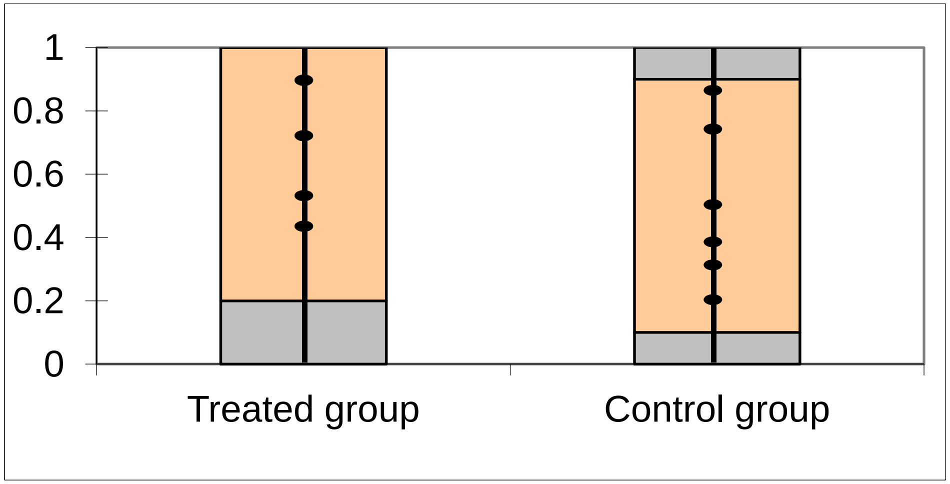
- Calculate the treatment effects: compare the outcomes *y* between the treated and control observations, after matching

$$y = \begin{cases} y_1 \text{ if } D = 1 \\ y_0 \text{ if } D = 0 \end{cases}$$

    - Counterfactual situation: compare the outcome of the treated observations with the outcome of the treated observations if they were not treated (find a close match using the control observations and use their outcome)

Matching methods explained

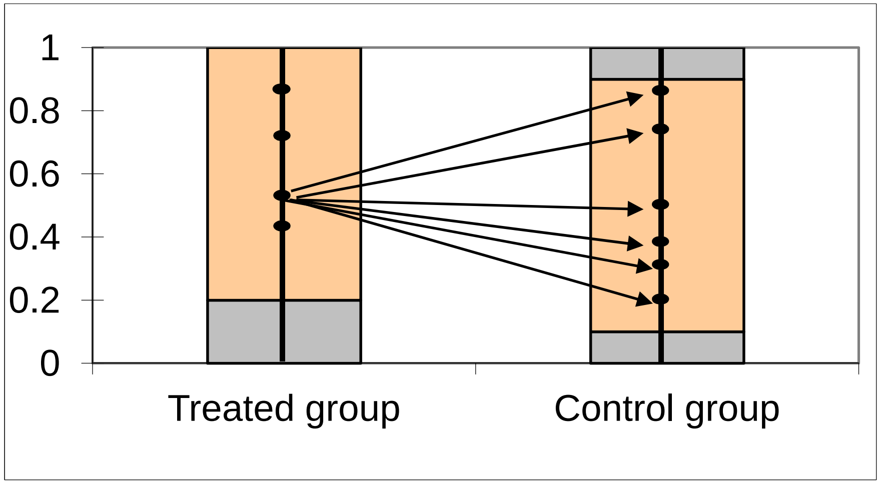
Propensity scores for treated and control groups



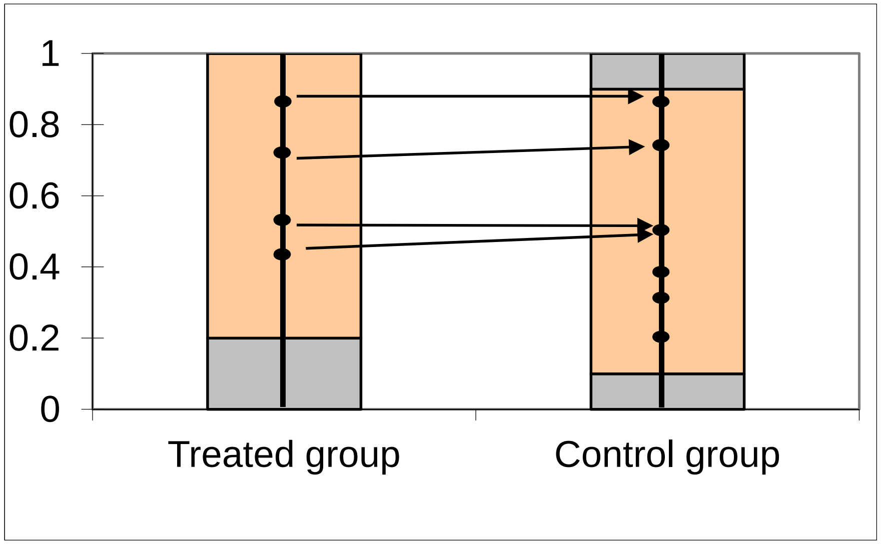
Matching methods: for each treated observation *i*, we need to find matches of control observation(s) *j* with similar characteristics.

- Matching with or without replacement
    - Matching without replacement - each control observation is used no more than one time as a match for a treated observation.
    - Matching with replacement – each control observation can be used as a match to several treated observations.

Kernel matching



Nearest neighbor matching



Nearest neighbor matching

- For each treated observation *i*, select a control observation *j* that has the closest *x*.

$$\min \parallel p_i - p_j \parallel$$

Radius matching

- Each treated observation *i* is matched with control observations *j* that fall within a specified radius.

$$\parallel p_i - p_j \parallel < r$$

Kernel matching

- Each treated observation *i* is matched with several control observations, with weights inversely proportional to the distance between treated and control observations.
- With matching based on propensity scores, the weights are defined as:

$$w(i,j) = \frac{K(\frac{p_j - p_i}{h})}{\sum_{j=1}^{n_0} K(\frac{p_j - p_i}{h})}$$

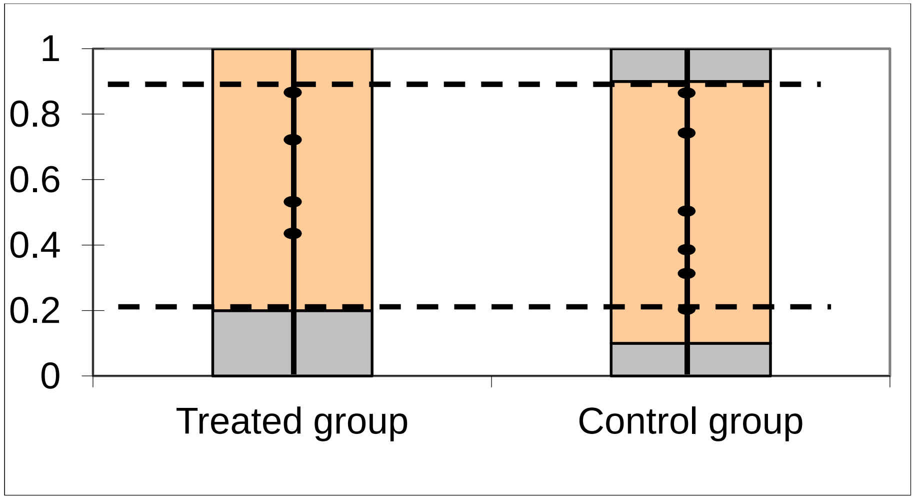
Here *h* is the bandwidth parameter.

Stratification or interval matching

- Compare the outcomes within intervals/blocks of propensity scores.

Matching with common support

- Restrict matching only based on the common range of propensity scores



Treatment effects

Average treatment effect (ATE)

- *ATE* is the difference between the outcomes of treated and control observations.

$$\Delta = y_1 - y_0$$

$$ATE = E(\Delta) = E(y_1|x, D = 1) - E(y_0|x, D = 0)$$

- A simple t-test between the outcomes for the treated and control groups.
- ATE is fine for random experiments but in observational studies, it may be biased if treated and control observations are not similar.

Average treatment effect on the treated (ATET)

- *ATET* is the difference between the outcomes of treated and the outcomes of the treated observations if they had not been treated.

$$ATET = E(\Delta|D = 1) = E(y_1|x, D = 1) - E(y_0|x, D = 1)$$

- The second term is a counterfactual so it is not observable and needs to be estimated.

Propensity score method

- After matching on propensity scores, we can compare the outcomes of treated and control observations.

$$ATET = E(\Delta|p(x), D = 1) = E(y_1|p(x), D = 1) - E(y_0|p(x), D = 0)$$

Empirical estimation

- Each treated observation *i* is matched *j* control observations and their outcomes y_0 are weighed by *w*.

$$ATET = \frac{1}{n_1} \sum_{i \in \{D=1\}} \left[y_{1,i} - \sum_j w(i,j) y_{0,j} \right]$$

Assumptions

- Partial equilibrium character (no general equilibrium effects)
    - Treatment does not indirectly affect the control observations.

Conditional independence assumption

- For random experiments, the outcomes are independent of treatment.

$$y_0, y_1 \perp D$$

- For observational studies, the outcomes are independent of treatment, conditional on x.

$$y_0, y_1 \perp D|x$$

- We need treatment assignment that ignores the outcomes.
- The treatment variable needs to be exogenous.

Unconfoundedness assumption

- Conditional independence of the control group outcome and treatment.
- Weaker assumption than the conditional independence assumption.

$$y_0 \perp D|x$$

Matching or overlap assumption

- For each value of x, there are both treated and control observations.
- For each treated observation, there is a matched control observation with similar x.

$$0 < prob(D = 1|x) < 1$$

Balancing condition

- Assignment to treatment is independent of the x characteristics, given the same propensity score.

$$D \perp x|p(x)$$

- The balancing condition is testable.

Difference-in-differences model

- The difference-in-differences model is applied when panel data on outcomes are available before (*b*) and after (*a*) the experiment occurs.
- The difference-in-differences model is an improvement over the one-period model.
- The difference-in-differences average treatment effect on the treated is specified as:

$$ATET = E(\Delta_a - \Delta_b | D = 1) = E((y_{1a} - y_{0a}) - (y_{1b} - y_{0b}) | x, D = 1) =$$

$$= E(y_{1a} - y_{1b} | x, D = 1) - E(y_{0a} - y_{0b} | x, D = 1)$$

- The first term refers to the differences in outcomes before and after the treatment for the treated group. This term may be biased if there are time trends. The second term uses the differences in outcomes for the control group to eliminate this bias.
- To apply the difference-in-differences model: instead of the outcomes for the treated and control groups, we use the differences in outcomes after the treatment and before the treatment. The rest of the analysis is the same.

Principal Component Analysis
and
Factor Analysis

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Principal Component Analysis and Factor Analysis

- PCA and factor analysis overview
- PCA methodology
- Component/factor retention
- Component/factor rotation (orthogonal vs. oblique)
- When to use principal component analysis
- Exploratory factor analysis

Principal Component Analysis and Factor Analysis

- Principal component analysis (PCA) and factor analysis are data reduction methods used to re-express multivariate data with fewer dimensions.
- The goal of these methods is to re-orient the data so that a multitude of original variables can be summarized with relatively few “factors” or “components” that capture the maximum possible information (variation) from the original variables.
- PCA is also useful in identifying patterns of association across variables.
- Factor analysis and principal component analysis are similar methods used for reduction of multivariate data; the difference between them is that factor analysis assumes the existence of a few common factors driving the variation in the data while principal component analysis does not make such an assumption.

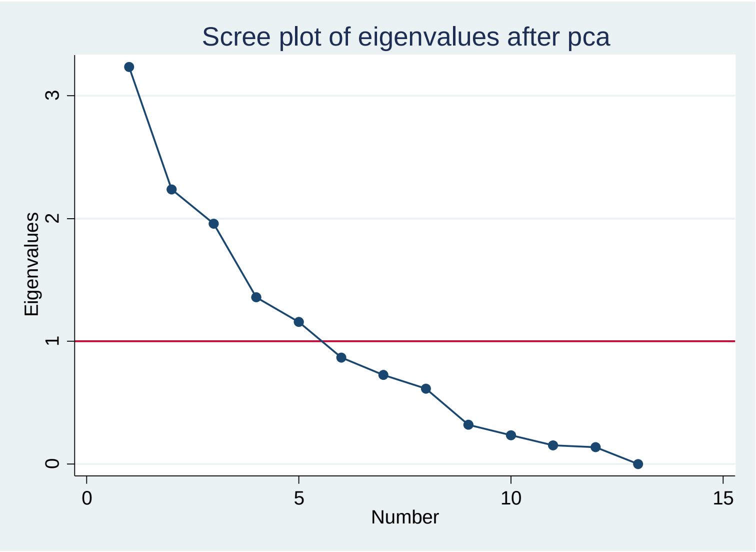
PCA Methodology

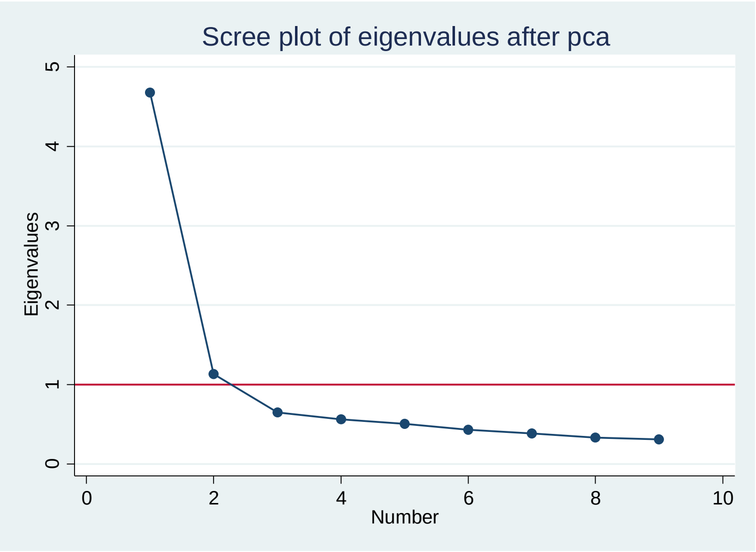
- The goal of PCA is to find components $z = [z_1 , z_2, \dots , z_p]$, which are a linear combination $u = [u_1 , u_2, \dots , u_p]'$ of the original variables $x = [x_1 , x_2, \dots , x_p]$ that achieve maximum variance.
- The first component z_1 is given by the linear combination of the original variables *x* and accounts for maximum possible variance. The second component captures most information not captured by the first component and is also uncorrelated with the first component.
- PCA seeks to maximize the variance so it is sensitive to scale differences in the variables. It is best to standardize the data and work with correlations rather than covariance among the original variables.
- PCA maximizes the variance of the elements of $z = xu$, such that $u'u = 1$.
- The solution is obtained by performing an eigenvalue decomposition of the correlation matrix, by finding the principal axes of the shape formed by the scatter plot of the data. The eigenvectors represent the direction of one of these principal axes.
- Solving the equation $(R - \lambda I)u = 0$, where *R* is the sample correlation matrix of the original variables x, λ is the eigenvalue, and *u* is the eigenvector.
- The eigenvalues λ are the variances of the associated components/factors *z*. The diagonal covariance matrix of the components is denoted as $D = diag(\lambda)$.

- The proportion of the variance in each original variable x_i accounted for by the first *c* factors is given by the sum of the squared factor loadings; that is, $\sum_{k=1}^{c} f_{ik}^{2}$. When *c*=*p* (all components are retained), $\sum_{k=1}^{c} f_{ik}^{2}$=1 (all variation in the data are explained).
- Factor loadings are the correlations between the original variables *x* and the components/factors z, denoted as $F = cor(x, z) = uD^{1/2}$.
  - Because the factor loadings matrix shows the correlation between the factors and the original variables, typically the factors are named after the set of variables they are most correlated with.
  - The components can also be "rotated" to simplify the structure of the loadings matrix and the interpretations of the results.

Factor Retention

- Since principal components analysis and factor analysis are data reduction methods, there is a need to retain an appropriate number of factors based on the trade-off between simplicity (retaining as few as possible factors) and completeness (explaining most of the variation in the data).
- The Kaiser's rule recommends retaining only factors with eigenvalues λ exceeding unity. Intuitively, this rule means that any retained factor z should account for at least as much variation as any of the original variables x.
- In practice, the scree plot of the eigenvalues is examined to determine whether there is a "break" in the plot with the remaining factors explaining considerably less variation.





Factor Rotation

- The factor loadings matrix is usually “rotated” or re-oriented in order to make most factor loadings on any specific factor small while only a few factor loadings large in absolute value.
- This simple structure allows factors to be easily interpreted as the clusters of variables that are highly correlated with a particular factor. The goal is to find clusters of variables that to a large extent define only one factor.

Orthogonal rotation – preserves the perpendicularity of the axes (rotated components/factors remain uncorrelated)

- Varimax rotation –preserves simple structure by focusing on the columns of the factor loading matrix. The Kaiser’s varimax rotation is an orthogonal rotation (preserving the independence of the factors) aiming to maximize the squared loadings variance across variables summed over all factors.
- Quartimax rotation – preserves simple structure by focusing on the rows of the factor loading matrix

Oblique rotation – allows for correlation between the rotated factors. The purpose is to align the factor axes as closely as possible to the groups of the original variables. The goal is to facilitate the interpretation of the results.

- Promax rotation

When to use Principal Component Analysis?

- Principal components analysis is undertaken in cases when there is a sufficient correlation among the original variables to warrant the factor/component representation.
- The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy takes values between 0 and 1, with small values indicating that overall the variables have little in common to warrant a principal components analysis and values above 0.5 are considered satisfactory for a principal components analysis.
- Bartlett’s sphericity test examines whether the correlation matrix should be factored, i.e. the data are not independent. It is a chi-square test with a test statistic that is a function of the determinant of the correlation matrix of the variables.

Exploratory Factor Analysis

- Common factor model – observed variance in each measure is attributable to a relatively small number of common factors and a single specific factor (unrelated to other factors in the model).

$$X_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \cdots \lambda_{ic}\xi_c + \delta_i$$

    - The common factors ξ contribute to the variation in all variables *X*.
    - The specific factor δ can be thought of as the error term.

- Factor analysis is appropriate when there is a "latent trait" or "unobservable characteristics."
- The factor scores can be obtained from the analysis of dependence.
- Factor analysis is used with survey questions about attitudes – the goal is to identify common factors capturing the variance from these questions and which can also be used as factor scores.

- Assumptions to determine a solution to the common factor model:
    - The common factors are uncorrelated with each other.
    - The specific factors are uncorrelated with each other.
    - The common factors and specific factors are uncorrelated with each other.

- The communality is the proportion of variance in *X* attributable to the common factors

$$h_i^2 = \sum_k \lambda_{ik}^2 = 1 - \theta_{ii}^2$$

  where $\theta_{ii}^2 = var(\delta_i)$ is the factor uniqueness.
- The solution to the common factor model is determined by orienting the first factor so that it captures the greatest possible variance, then the second factor is obtained so that it captures the greatest possible variance but is uncorrelated with the first factor.

- The correlations between the X variables and the Ξ factors are called factor loadings Λ.
- The factor scores present the positions of the observations in the common factor space. The factor score coefficients are given by

$$B = R^{-1}\Lambda_c$$

  where R is the correlation matrix.
- The factor scores are calculated as:

$$\Xi = X_s B$$

- These factor scores are included in the data and can be used instead of the original variables.

Instrumental Variables

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Instrumental Variables Overview

- Endogeneity examples
- Endogeneity definitions
- Instrumental variables set up
- The two stage least squares (2SLS) estimation procedure
- Identification issues
- Endogeneity tests
- Weak instrumental variables
- Systems of equations (2SLS and 3SLS)

Instrumental Variables

Endogeneity examples

- Wages and education jointly depend on ability which is not directly observable. We can use available test results to proxy for ability.
- Consumption and income are both determined by macroeconomic factors. We can use investments to control for endogeneity.

Causes of endogeneity

- The explanatory variables are measured with errors
- Reverse causality (the explanatory variable is caused by the dependent variable)

Endogeneity definitions

A regressor is endogenous when it is correlated with the error term.
Example: y is earnings, x is years of schooling, u is error term (including ability), z is proximity to college.

Exogeneity: regressors *x* and the error term *u* are independent causes of the dependent variable *y*.

$$\begin{tikzcd} x \arrow[r] & y \\ u \arrow[ur] & \end{tikzcd}$$

Endogeneity: the error *u* is affecting the regressors *x* and therefore indirectly affecting *y*.

$$\begin{tikzcd} x \arrow[r] & y \\ u \arrow[u] \arrow[ur] & \end{tikzcd}$$

Instrumental variables: instruments *z* are associated with *x* but not with the error term *u*.

$$\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & u \arrow[u] \arrow[ur] & \end{tikzcd}$$

Requirements for instruments *z*:

- *z* is correlated with the regressors *x*, $\mathrm{E}[z'x] \neq 0$ (*z* predicts or causes *x*),
- *z* is uncorrelated with the error term *u*, $\mathrm{E}[z'u] = 0$ (*z* is not endogenous),
- *z* is not a direct cause of the dependent variable *y*, $\mathrm{cov}[y, z|x] = 0$ (*z* is not in the *y* equation).

Instrumental variables set up

- Consider the linear model: $y = x\beta + u$
- Endogeneity is when one or more explanatory variables are correlated with the error term: $\mathrm{E}[x|u] = \mathrm{cov}(x'u) \neq 0$.
- The estimated coefficients from the OLS estimation are biased:
$$b = \beta + (x'x)^{-1}x'u,\ \mathrm{E}[b] \neq \beta.$$

- We re-write the model as the following structural equation:
$$y_1 = y_2'\beta_1 + x_1'\beta_2 + u$$

  where y_1 is the dependent variable, y_2 is the endogenous variable, and x_1 are the exogenous variables.

- The structural equation model involves a combined set $x = [y_2, x_1]$ of both endogenous and exogenous variables.
- We need to find a set of instrument $z = [x_1, x_2]$ of only exogenous variables, where x_1 is instrument for itself and x_2 is instrument for y_2.

The two stage least squares (2SLS) estimation procedure

- The 2SLS procedure replaces the endogenous variable with predicted values of this endogenous variable when regressed on instruments.

1. Estimate the first stage (reduced form) equation with only exogenous regressors.

$$y_2 = x_1^{'}\gamma_1 + x_2^{'}\gamma_2 + e$$

2. Calculate the predicted values $\hat{y}_2$and substitute them in the structural equation model.

$$y_1 = \hat{y}_2^{'}\beta_1 + x_1^{'}\beta_2 + u$$

Identification issues

- Order condition: The number of omitted instrumental variables must be at least as large as the number of endogenous regressor.
- Rank condition: The matrices $z'x$ must have a full rank in order to be inverted.

Just-identified model

- An IV model is just identified if there is one instrument $x_2$for each endogenous variable y_2.

$$b_{IV} = (z'x)^{-1}z'y = (z'x)^{-1}z'(x\beta + u) = \beta + (z'x)^{-1}z'u$$

- This estimator is unbiased.

Under-identified model

- An IV model is under-identified if there are fewer instruments $x_2$than endogenous variables y_2.
- The under-identified model has an infinite number of solutions and therefore no consistent estimator exists.

Over-identified model

- An IV model is over-identified if there are more instruments than endogenous variables.
- There are two efficient estimators that can be used:

- The two stage least squares (2SLS) (best if the error term is iid and homoskedastic):

$$b_{2sls} = [x'z(z'z)^{-1}z'x]^{-1}x'z(z'z)^{-1}z'y$$

- The generalized method of moments (GMM):

$$b_{GMM} = (x'zwz'x)^{-1}x'zwz'y$$

If $w = (z'z)^{-1}$, then this is the 2SLS estimate.
Usually $w = \hat{S}^{-1}$, where $\hat{S}$ is the estimated variance of *z'u.*
This estimator is optimal in presence of heteroscedasticity.

Instrumental variables tests

Hausman test for endogeneity

- The Hausman test checks if a regressor is exogenous or endogenous.
- The Hausman test compares the OLS and IV estimates to check for significant differences.
  - If there are significant differences, then the regressor is endogenous.
  - If there are no significant differences, then the regressor is exogenous.

Durbin-Wu-Hausman test for exogenous regressors

- The Durbin-Wu-Hausman test is a procedure that checks whether $\mathrm{E}[\mathrm{x}|e] = \mathrm{cov}(\mathrm{x}e) \neq 0$.
- Estimate the first-stage model: $y_2 = x_1^{'}\gamma_1 + x_2^{'}\gamma_2 + u$
- Include the residuals ($\hat{u}$) from the first-stage regression in the structural equation regression: $y_1 = y_2^{'}\beta_1 + x_1^{'}\beta_2 + \hat{u}\rho + e$
  - If the coefficient on the residuals from the first-stage regression ρ is not significantly different from zero then the regressors are exogenous.
  - If the coefficient ρ is significantly different from zero then the regressors are endogenous.

Tests for overidentifying restrictions

- Estimate model using GMM and form a test statistic:

$$Q(\beta) = (1/N)(y - x\beta)'z(S^{-1})(1/N)z'(y - x\beta)$$

- It is distributed as chi-square with degrees of freedom of the number of overidentifying restrictions.
- Rejection of null hypothesis – at least one instrument is not valid.

Weak Instrumental Variables

A weak instrument has a low correlation with the endogenous variable.

Tests for weak instruments

- In a case of one endogenous regressor and one instrument, a low correlation between instrument and the endogenous variable would indicate a weak instrument.
- When several instruments are used for one endogenous variable, the weakness of the instruments can be measured by the partial R^2 and partial F-statistic from the first stage regression.
    - The instrument is weak if the partial F-statistic testing the joint significance of the coefficients of the instruments ($\gamma_2 = 0$) is less than 10.

Consequences of weak instruments

- A weak instrument will undermine the precision of the estimator.

$$V(\hat{\beta}_{IV}) = V(\hat{\beta}_{OLS})/r_{xz}^2$$

- The IV estimator is asymptotically consistent but biased toward OLS estimator in finite-sample. The size of the bias is positively related to the weakness of the instrument(s) and inversely related with the size of the sample.

Instrumental Variables and Simultaneous Systems of Equations

Simultaneous systems of equations with two endogenous variables

- The system of structural equations is:

$$y_1 = y_2'\beta_1 + z_1'\gamma_1 + u_1$$
$$y_2 = y_1'\beta_2 + z_2'\gamma_2 + u_2$$

- There are endogenous variables as independent variables in both equations.
- The reduced form equation is:

$$y = z'\Gamma + u$$

The two stage least squares (2SLS) or three stage least squares (3SLS) procedure:

1. Estimate the reduced form equation by OLS regression and obtain $\hat{y}$.
2. Use the estimates $\hat{y}$ from the first stage to estimate the structural equations:

$$y_1 = \hat{y}_2'\beta_1 + z_1'\gamma_1 + u_1$$
$$y_2 = \hat{y}_1'\beta_2 + z_2'\gamma_2 + u_2$$

These estimates are the 2SLS estimates.

3. Use the 2SLS estimates to compute the 3SLS using the following estimator:

$$\hat{\beta}_{3SLS} = \{X'(\Sigma^{-1} \otimes I_N)X\}^{-1}\{X'(\Sigma^{-1} \otimes I_N)y\}$$

2SLS and 3SLS comparison

- 3SLS is more efficient than 2SLS because it uses cross-equation information.
- 3SLS is inconsistent if the error term is heteroscedastic.

Seemingly Unrelated Regressions (SUR)

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Seemingly Unrelated Regressions Overview

- Systems of equations examples
- SUR model
- SUR estimation
- Properties of the SUR model

Seemingly Unrelated Regressions (SUR)

Systems of equations examples

- Demand system for food items for individuals
- Expenditure system for several different types of expenditures
- Demand and supply models

Linear systems of equations

- Systems of equations include multiple equations instead of one equation.
    - Simultaneous Equation Models: Contain both endogenous and exogenous regressors.
    - Seemingly Unrelated Regression (SUR) Models: Contain only exogenous regressors.

SUR model

- SUR model is a system of linear equations with errors that are correlated across equations for a given individual but are uncorrelated across individuals.
- The model consists of *j*=1…*m* linear regression equations for *i*=1…*N* individuals. The *j*th equation for individual *i* is

$$y_{ij} = x'_{ij}\beta_j + u_{ij}$$

- With all observations stacked, the model for the *j*th equation can be written as

$$y_j = x'_j\beta_j + u_j$$

- We can stack the *m* equations into an SUR model:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ . & & \ddots & \vdots \\ 0 & & \cdots & X_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_m \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_m \end{bmatrix}$$

- The error terms are assumed to have zero mean and to be independent across individuals and homoskedastic.
- For a given individual, the errors are correlated across equations,

$$E\left(u_{ij}u_{ij'}\middle|X\right) = \sigma_{jj'}, and\ \sigma_{jj'} \neq 0\ where\ j \neq\ j'$$

- The error term u_j satisfies the following assumptions:
  - Mean of error term: $E\left(u_j \middle| X\right) = 0$
  - Variance of error term in equation *j*: $E\left(u_j u'_j \middle| X\right) = \sigma_{jj} I_N$
  - Covariance of error terms across equations *j* and *j'*: $E\left(u_j u'_{j'} \middle| X\right) = \sigma_{jj'} I_N$ where $j \neq \ j'$
  - Overall variance-covariance matrix: $\Omega = E\left(uu'\right) = \Sigma \otimes I_N$

SUR estimation

- OLS estimation for each equation yields a consistent estimator of β, but the optimal estimator is the GLS estimator.

$$\hat{\beta}_{GLS} = \left\{X'\left(\Sigma^{-1}\otimes I_N\right)X\right\}^{-1}\left\{X'\left(\Sigma^{-1}\otimes I_N\right)y\right\}$$

$$\text{with } Var\left(\hat{\beta}\right) = \left\{X'\left(\Sigma^{-1}\otimes I_N\right)X\right\}^{-1}.$$

- Two steps estimation:
    - First step: each equation is estimated by OLS, and the residuals from the *m* equations are used to estimate Σ; using $\widehat{u_j} = y_j - X_j\widehat{\beta_j}$, and $\hat{\sigma}_{jj'} = \frac{\widehat{u_j}'\widehat{u_{j'}}}{N}$
    - Second step: substitute $\widehat{\Sigma}$ for Σ of GLS estimator.

$$\hat{\beta}_{GLS} = \left\{X'\left(\widehat{\Sigma}^{-1}\otimes I_N\right)X\right\}^{-1}\left\{X'\left(\widehat{\Sigma}^{-1}\otimes I_N\right)y\right\}$$

- Testing cross-equation restrictions and imposing constraints:
  - We can test whether the coefficients are jointly significantly different from zero $\beta_j = \beta_{j'} = 0$ or whether the coefficients are significantly different from each other $\beta_j = \beta_{j'}$.
  - We can also impose a cross-equation restriction $\beta_j = \beta_{j'}$ and then estimate the SUR model.

Properties of the SUR model

- The SUR model is used to gain efficiency when the equations are only related through the error term.
- The parameters in the SUR model generally vary from equation to equation.
- Regressors may or may not vary from equation to equation depending on the model.
- The SUR estimates result in equation-by-equation OLS estimates when:
  - The errors are uncorrelated across equations, so Σ is diagonal.
  - Each of the equations contains exactly the same set of regressors, so $X_j = X_{j'}$.

Time Series ARIMA Models

Ani Katchova

© 2013 by Ani Katchova. All rights reserved.

Time Series Models Overview

- Time series examples
- White noise, autoregressive (AR), moving average (MA), and ARMA models
- Stationarity, detrending, differencing, and seasonality
- Autocorrelation function (ACF) and partial autocorrelation function (PACF)
- Dickey-Fuller tests
- The Box-Jenkins methodology for ARMA model selection

Time Series ARIMA Models

Time series examples

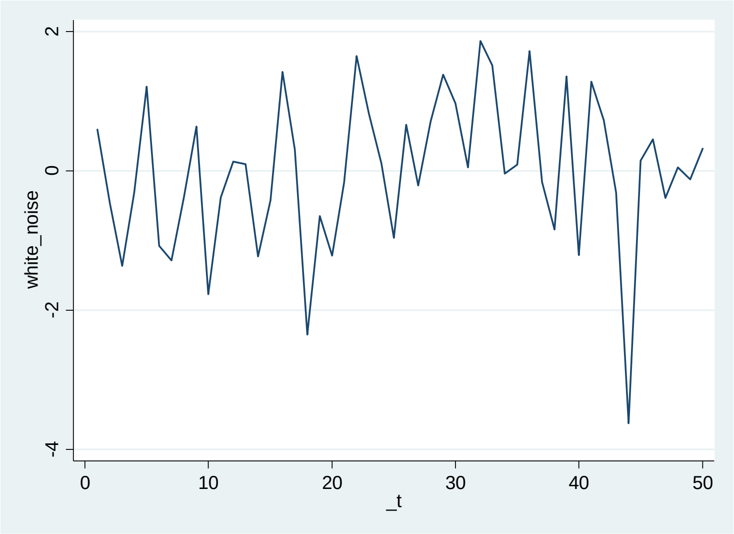
- Modeling relationships using data collected over time – prices, quantities, GDP, etc.
- Forecasting – predicting economic growth.
- Time series involves decomposition into a trend, seasonal, cyclical, and irregular component.

Problems ignoring lags

- Values of y_t are affected by the values of y in the past.
    - For example, the amount of money in your bank account in one month is related to the amount in your account in a previous month.
- Regression without lags fails to account for the relationships through time and overestimates the relationship between the dependent and independent variables.

White noise

- White noise describes the assumption that each element in a series is a random draw from a population with zero mean and constant variance.

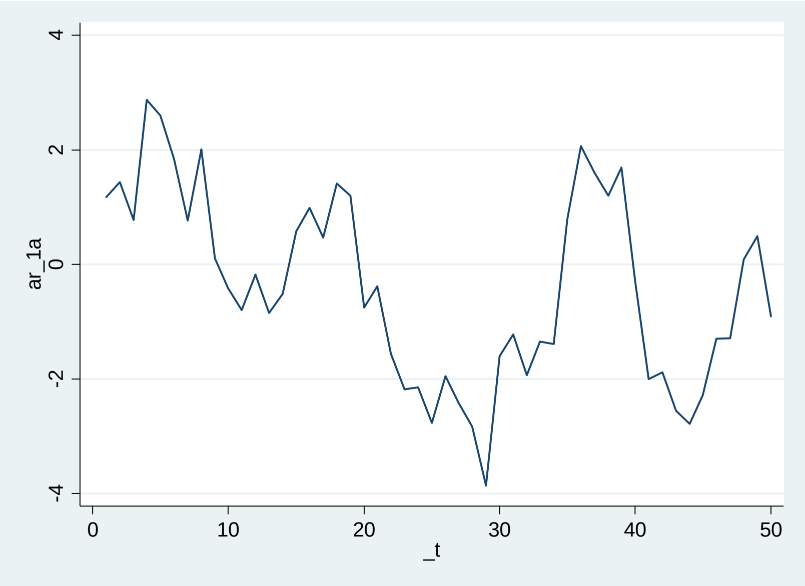


- Autoregressive (AR) and moving average (MA) models correct for violations of this white noise assumption.

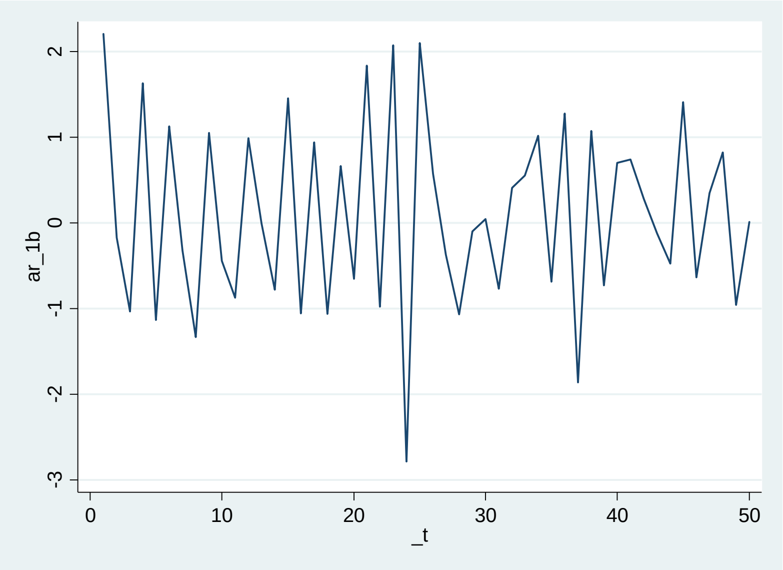
Autoregressive (AR) models

- Autoregressive (AR) models are models in which the value of a variable in one period is related to its values in previous periods.
- AR(p) is an autoregressive model with p lags: $y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$
  where μ is a constant and γ_p is the coefficient for the lagged variable in time *t-p*.
- AR(1) is expressed as: $y_t = \mu + \gamma y_{t-1} + \epsilon_t = \mu + \gamma(Ly_t) + \epsilon_t$ or $(1 - \gamma L)y_t = \mu + \epsilon_t$

AR(1) with $\gamma = 0.8$



AR(1) with $\gamma = -0.8$



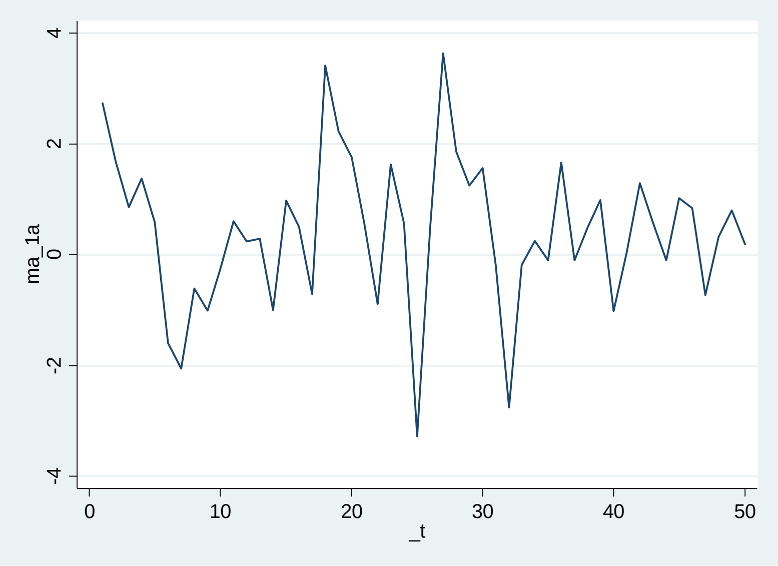
Moving average (MA) models

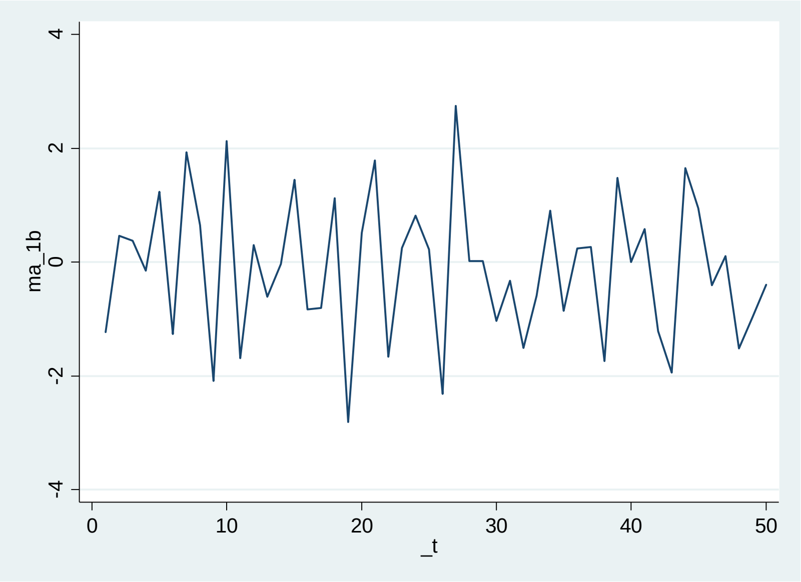
- Moving average (MA) models account for the possibility of a relationship between a variable and the residuals from previous periods.
- MA(q) is a moving average model with q lags: $y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$ where θ_q is the coefficient for the lagged error term in time *t*-*q*.

- MA(1) model is expressed as: $y_t = \mu + \epsilon_t + \theta \epsilon_{t-1}$
- Note: SAS (unlike Stata and R), model θ with a reverse sign.

MA(1) with $\theta = 0.7$

MA(1) with $\theta = -0.7$



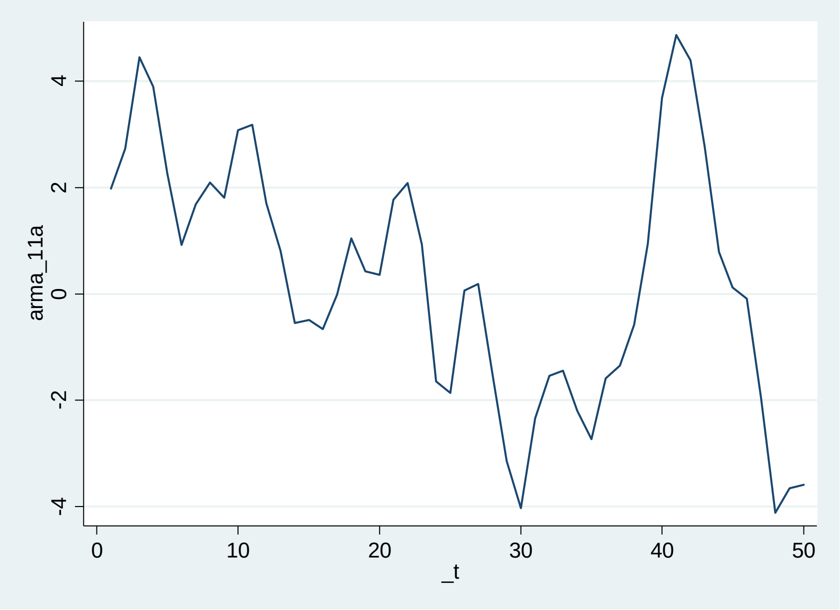


Autoregressive moving average (ARMA) models

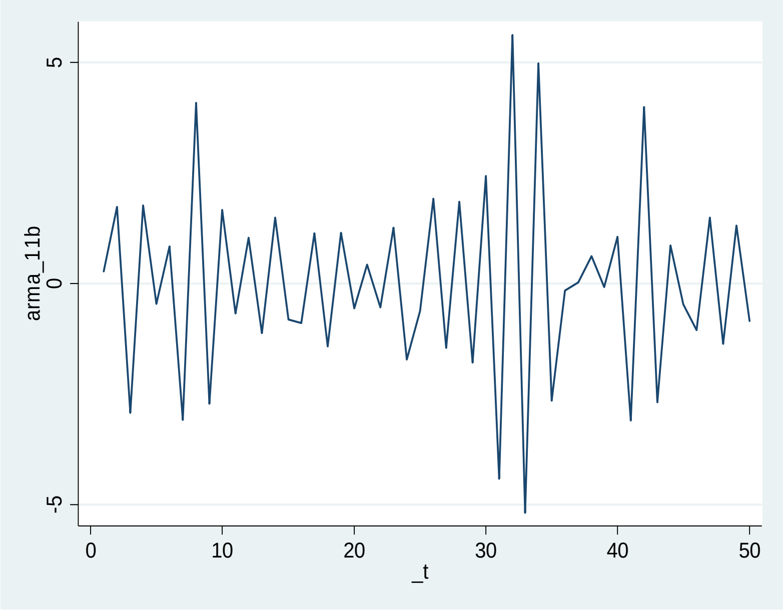
- Autoregressive moving average (ARMA) models combine both p autoregressive terms and q moving average terms, also called ARMA(p,q).

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

ARMA(1,1) with $\gamma = 0.8$ and $\theta = 0.7$



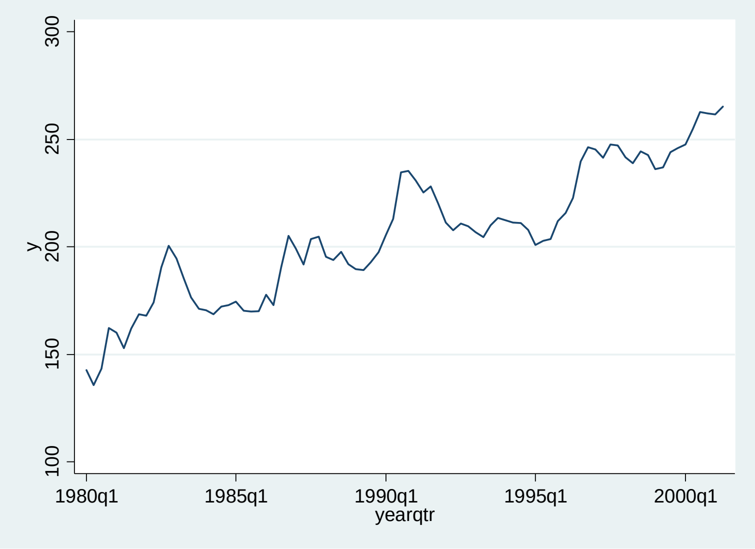
ARMA(1,1) with $\gamma = -0.8$ and $\theta = -0.7$



Stationarity

- Modeling an ARMA(p,q) process requires stationarity.
- A stationary process has a mean and variance that do not change over time and the process does not have trends.
- An AR(1) disturbance process: $u_t = \rho u_{t-1} + \epsilon_t$
- is stationary if $|\rho| < 1$ and ϵ_t is white noise.

- Example of a time-series variable, is it stationary?

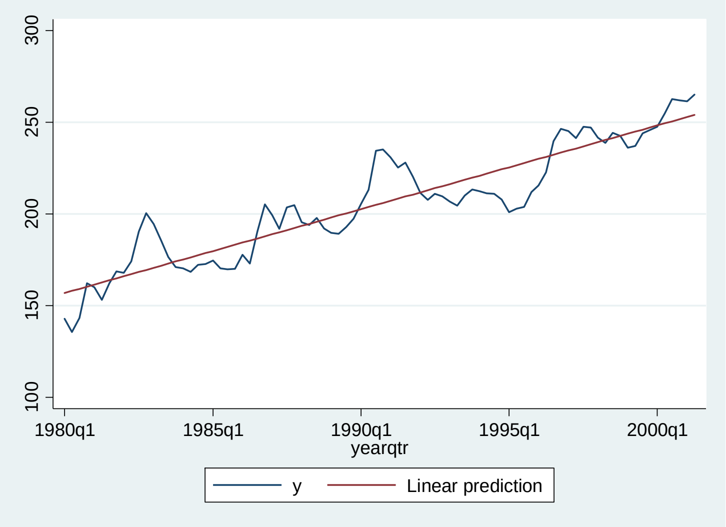


Detrending

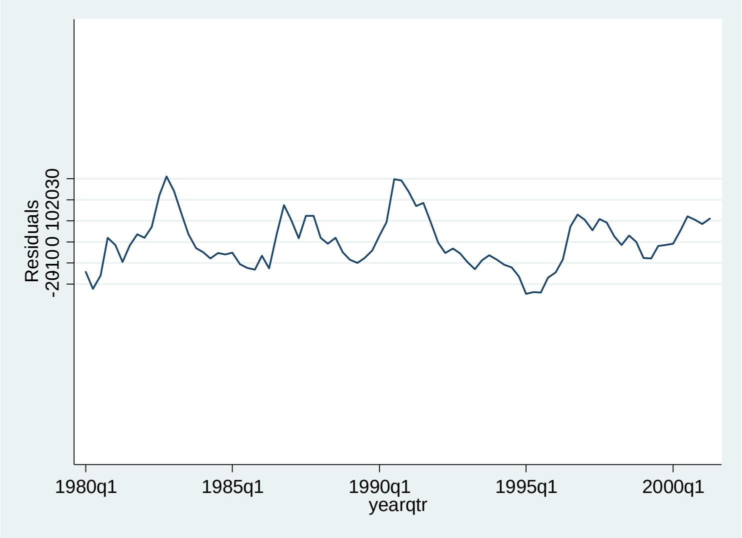
- A variable can be detrended by regressing the variable on a time trend and obtaining the residuals.

$$y_t = \mu + \beta t + \varepsilon_t$$

Variable y_t



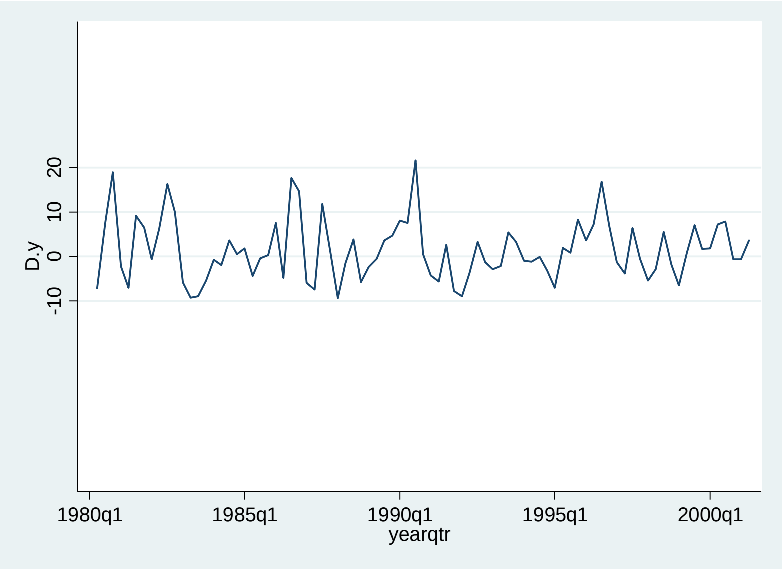
Detrended variable: $\hat{\varepsilon}_t = y_t - \hat{\mu} + \hat{\beta} t$



Differencing

- When a variable y_t is not stationary, a common solution is to use differenced variable: $\Delta y_t = y_t - y_{t-1}$, for first order differences.
- The variable y_t is integrated of order one, denoted $I(1)$, if taking a first difference produces a stationary process.
- ARIMA (p,d,q) denotes an ARMA model with p autoregressive lags, q moving average lags, a and difference in the order of d.

Differenced variable: $\Delta y_t = y_t - y_{t-1}$



Seasonality

- Seasonality is a particular type of autocorrelation pattern where patterns occur every “season,” like monthly, quarterly, etc.
- For example, quarterly data may have the same pattern in the same quarter from one year to the next.
- Seasonality must also be corrected before a time series model can be fitted.

Dickey-Fuller Test for Stationarity

Dickey-Fuller test

- Assume an AR(1) model. The model is non-stationary or a unit root is present if $|\rho| = 1$.

$$y_t = \rho y_{t-1} + e_t$$

$$y_t - y_{t-1} = \rho y_{t-1} - y_{t-1} + e_t$$

$$\Delta y_t = (\rho - 1) y_{t-1} + e_t = \gamma y_{t-1} + e_t$$

- We can estimate the above model and test for the significance of the γ coefficient.
  - If the null hypothesis is not rejected, $\gamma^* = 0$, then y_t is not stationary. Difference the variable and repeat the Dickey-Fuller test to see if the differenced variable is stationary.
  - If the null hypothesis is rejected, $\gamma^* > 0$, then y_t is stationary. Use the variable.
  - Note that non-significance is means stationarity.

Augmented Dickey-Fuller test

- In addition to the model above, a drift μ and additional lags of the dependent variable can be added.

$$\Delta y_t = \mu + \gamma^* y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \epsilon_t$$

- The augmented Dickey-Fuller test evaluates the null hypothesis that $\gamma^* = 0$. The model will be non-stationary if $\gamma^* = 0$.

Dickey-Fuller test with a time trend

- The model with a time trend:

$$\Delta y_t = \mu + \beta t + \gamma^* y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \epsilon_t$$

- Test the hypothesis that $\beta = 0$ and $\gamma^* = 0$. Again, the model will be non-stationary or will have a unit root present if $\gamma^* = 0$.

Autocorrelation Function (ACF) and Partial Autocorrelation Function (ACF)

Autocorrelation function (ACF)

- ACF is the proportion of the autocovariance of y_t and y_{t-k} to the variance of a dependent variable y_t

$$ACF(k) = \rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\mathrm{Var}(y_t)}$$

- The autocorrelation function ACF(k) gives the gross correlation between y_t and y_{t-k}.
- For an AR(1) model, the ACF is $ACF(k) = \rho_k = \gamma^k$. We say that this function tails off.

Partial autocorrelation function (PACF)

- PACF is the simple correlation between y_t and y_{t-k} minus the part explained by the intervening lags

$$\rho_k^* = \mathrm{Corr}[y_t - E^*(y_t | y_{t-1}, \dots, y_{t-k+1}), y_{t-k})]$$

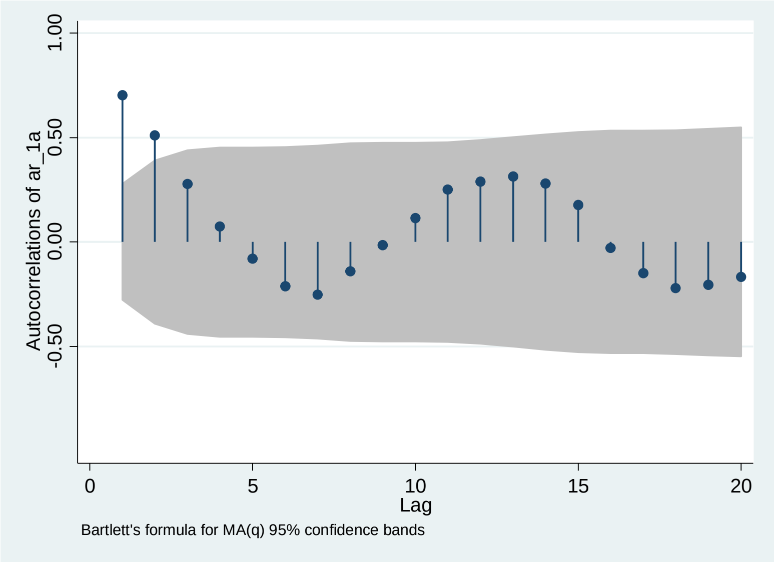
where $E^*(y_t | y_{t-1}, \dots, y_{t-k+1})$ is the minimum mean-squared error predictor of y_t by $y_{t-1}, \dots, y_{t-k+1}$.

- For an AR(1) model, the PACF is γ for the first lag and then cuts off.

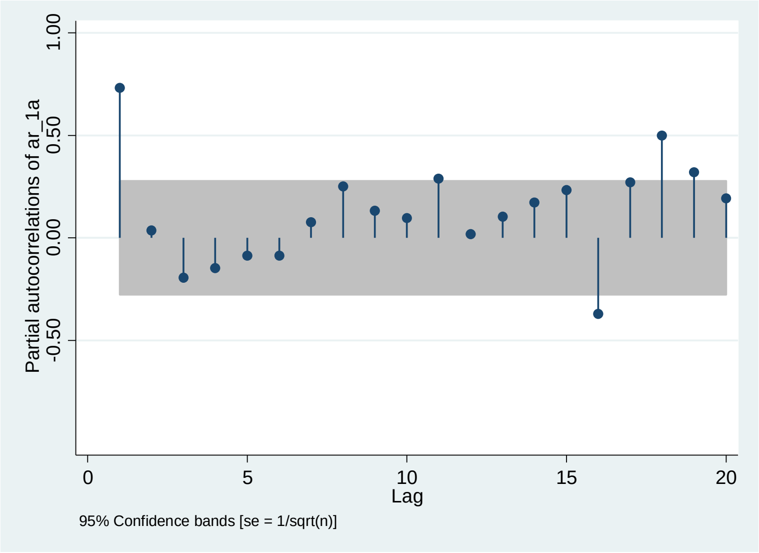
ACF and PACF properties

	AR(*p*)	MA(*q*)	ARMA(*p*,*q*)
ACF	Tails off	Cuts off after lag *q*	Tails off
PACF	Cuts off after lag *p*	Tails off	Tails off

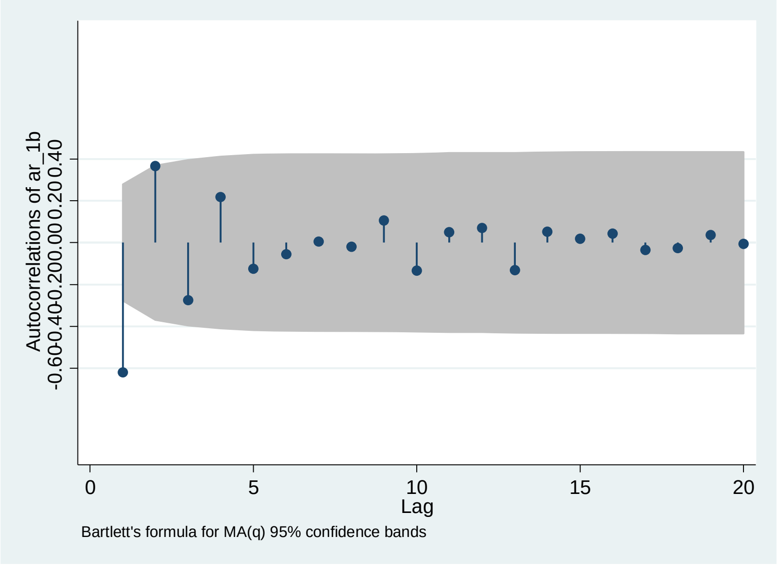
ACF of AR(1) with coefficient 0.8



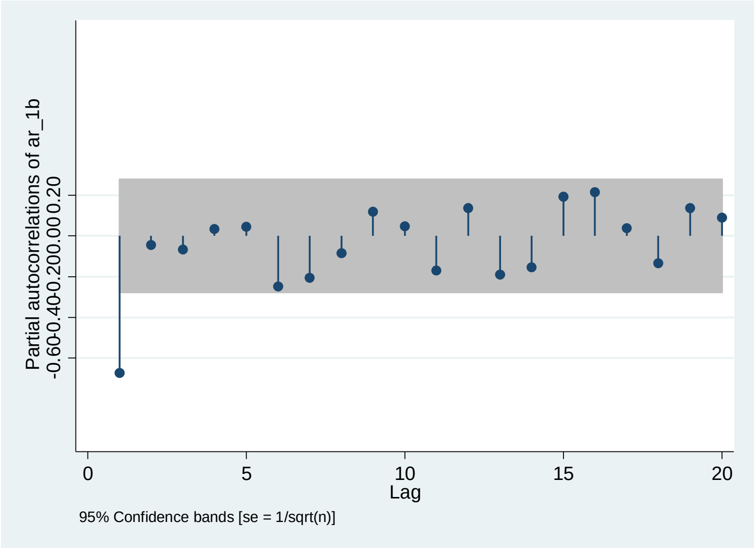
PACF of AR(1) with coefficient of 0.8



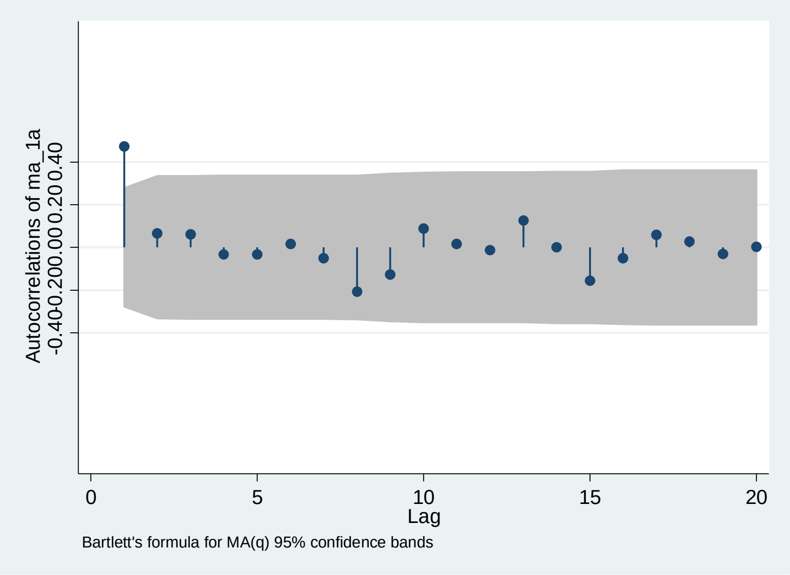
ACF of AR(1) with coefficient -0.8



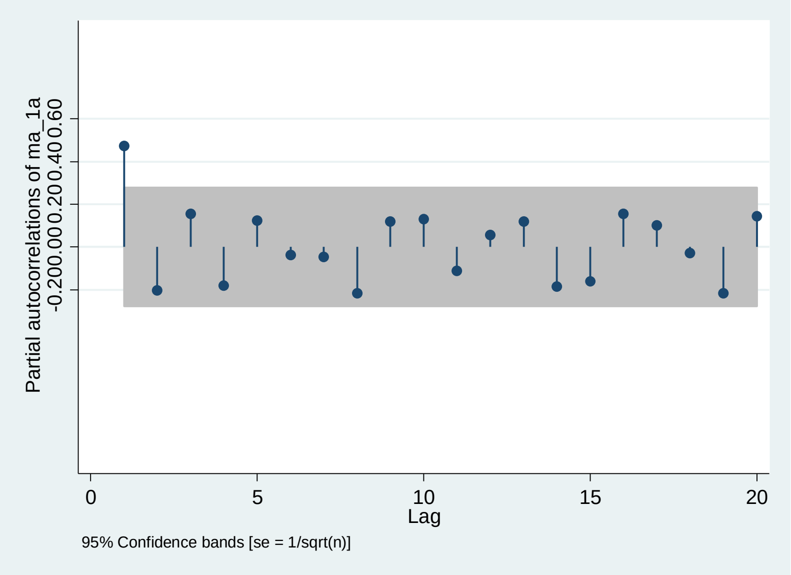
PACF of AR(1) with coefficient of -0.8



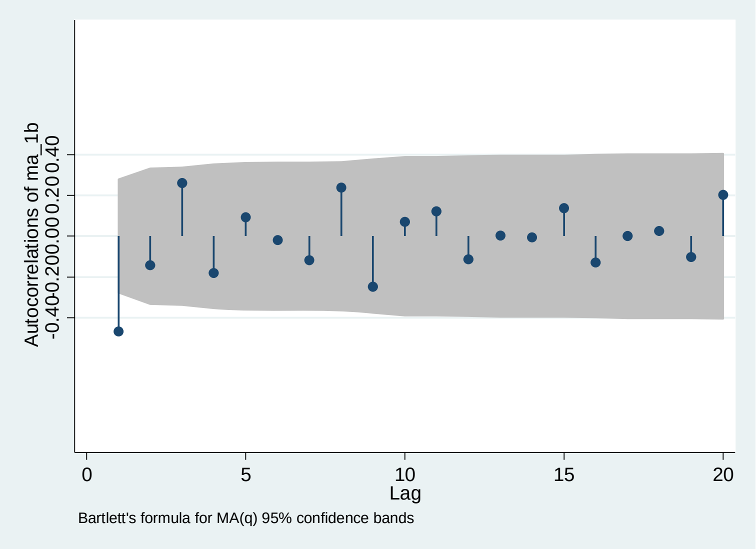
ACF of MA(1) with coefficient of 0.7



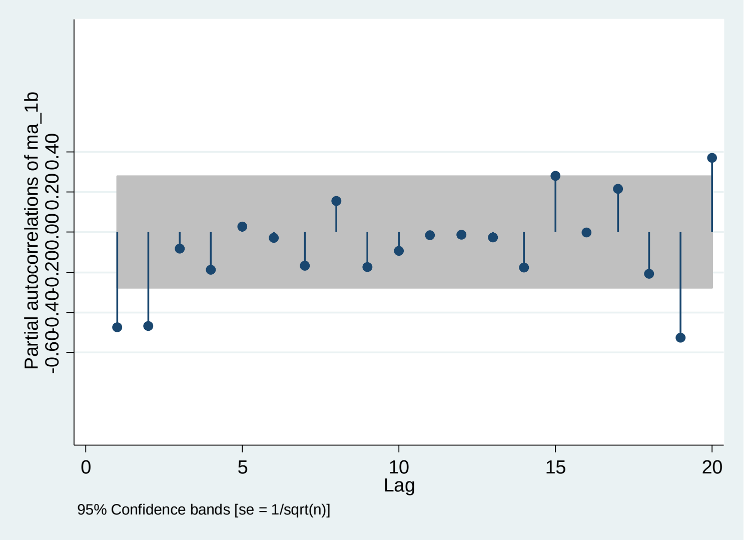
PACF of MA(1) with coefficient of 0.7



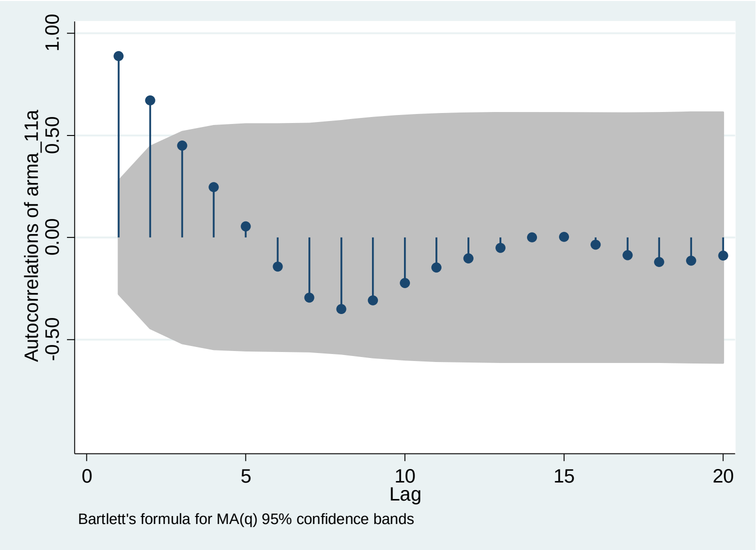
ACF of MA(1) with coefficient of -0.7



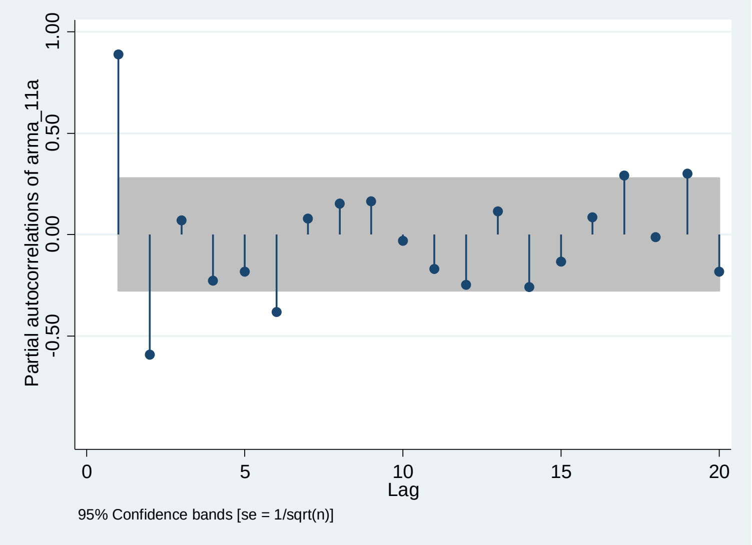
PACF of MA(1) with coefficient of -0.7



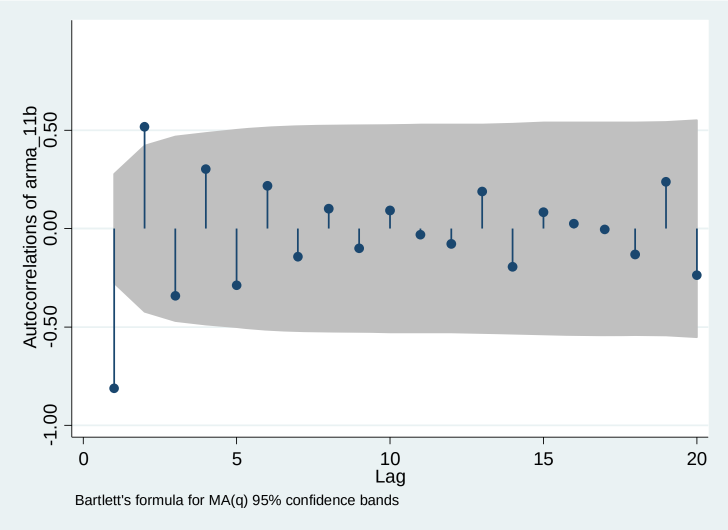
ACF of ARMA(1,1) with coeff 0.8 and 0.7



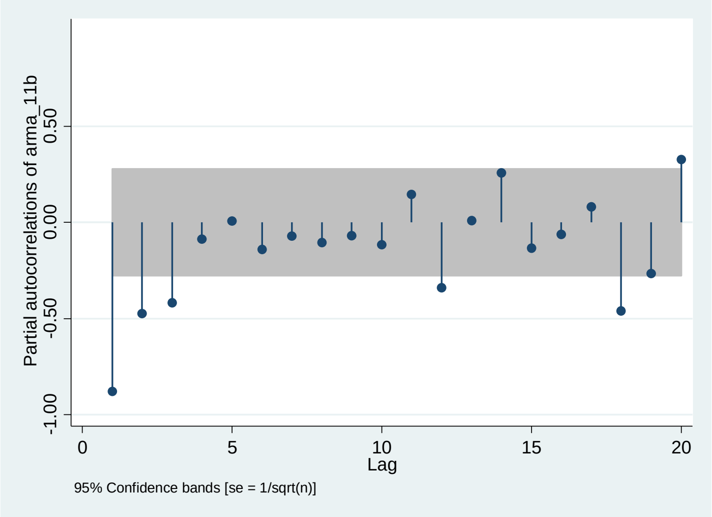
PACF of ARMA(1,1) with coeff 0.8 and 0.7



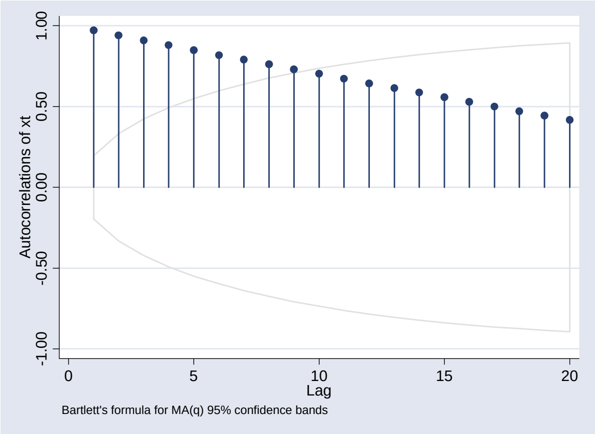
ACF of ARMA(1,1) with coeff -0.8 and -0.7



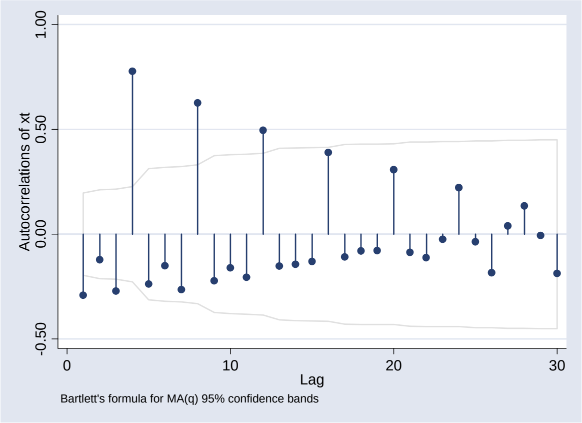
PACF of ARMA(1,1) with coeff -0.8 and -0.7



ACF of non-stationary series - The ACF shows a slow decaying positive ACF.



ACF with seasonal lag (4) – ACF shows spikes every 4 lags.



Diagnostics for ARIMA Models

Testing for white noise

- The Box-Pierce statistic is the following: $Q = T\sum_{k=1}^{P}\rho_k^2$
- The Ljung-Box statistic: $Q' = T(T+2)\sum_{k=1}^{P}\frac{\rho_k^2}{T-k}$

  where ρ_k is the sample autocorrelation at lag k.
- Under the null hypothesis that the series is white noise (data are independently distributed), Q has a limiting χ^2 distribution with *p* degrees of freedom.

Goodness of fit

- Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two measures goodness of fit. They measure the trade-off between model fit and complexity of the model.

$$\mathrm{AIC} = -2\ln(L) + 2k$$

$$\mathrm{BIC} = -2\ln(L) + \ln(N)k$$

where L is the value of the likelihood function evaluated at the parameter estimates, N is the number of observations, and k is the number of estimated parameters.

- A lower AIC or BIC value indicates a better fit (more parsimonious model).

The Box-Jenkins Methodology for ARIMA Model Selection

Identification step

- Examine the time plot of the series.
  - Identify outliers, missing values, and structural breaks in the data.
  - Non-stationary variables may have a pronounced trend or have changing variance.
  - Transform the data if needed. Use logs, differencing, or detrending.
    - Using logs works if the variability of data increases over time.
    - Differencing the data can remove trends. But over-differencing may introduce dependence when none exists.
- Examine the autocorrelation function (ACF) and partial autocorrelation function (PACF).
  - Compare the sample ACF and PACF to those of various theoretical ARMA models. Use properties of ACF and PACF as a guide to estimate plausible models and select appropriate p, d, and q.
  - With empirical data, several models may need to be estimated.
  - Differencing may be needed if there is a slow decay in the ACF.

Estimation step

- Estimate ARMA models and examine the various coefficients.
- The goal is to select a stationary and parsimonious model that has significant coefficients and a good fit.

Diagnostic checking step

- If the model fits well, then the residuals from the model should resemble a while noise process.
    - Check for normality looking at a histogram of the residuals or by using a quantile-quantile (Q-Q) plot.
    - Check for independence by examining the ACF and PACF of the residuals, which should look like a white noise.
    - The Ljung-Box-Pierce statistic performs a test of the magnitude of the autocorrelations of the correlations as a group.
    - Examine goodness of fit using the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Use most parsimonious model with lowest AIC and/or BIC.