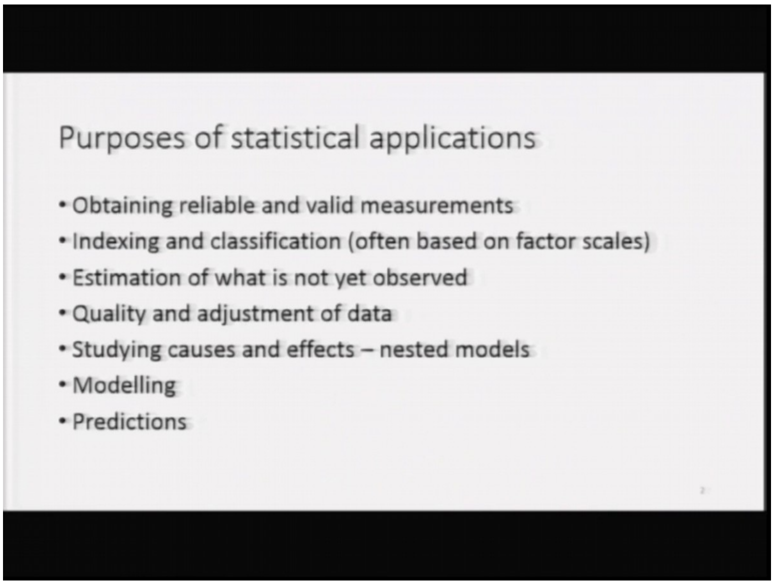
Population Studies
Prof. Arun Kumar Sharma
Department of Humanities and Social Sciences
Indian Institute of Technology, Kanpur

Lecture - 16
Statistical Techniques in Population Studies - II

Dear students, we are talking about Mathematical Models and Statistical Techniques in Population Studies. In the last lecture I gave you some examples of application of mathematical and statistical models and I showed that these models are used for estimating the unknown, for predictive purposes, for evaluation of quality of certain types of data, and for explanatory purposes and I gave you some examples, I did not actually discuss models in detail; but I just gave you examples of what kind of questions models can answer.

I will use the same strategy in this lecture and show you some more models of analysis in population studies and for answering what kind of questions they have been used; those who have right background, they can go through these references and learn more about the models.

(Refer Slide Time: 01:27)



Purposes of statistical applications

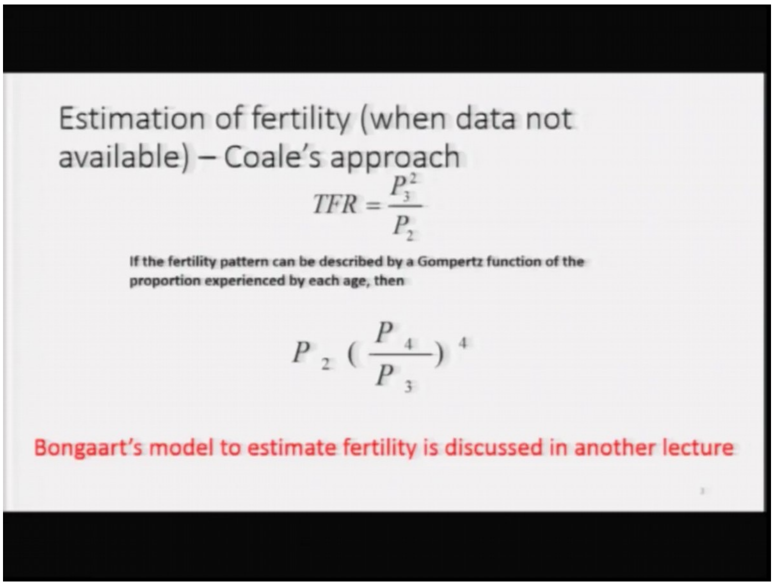
- Obtaining reliable and valid measurements
- Indexing and classification (often based on factor scales)
- Estimation of what is not yet observed
- Quality and adjustment of data
- Studying causes and effects – nested models
- Modelling
- Predictions

The purpose of statistical applications and that includes mathematical modeling is; obtaining reliable and valid measurements, indexing and classification often based on factor scales. If you look at say wealth index of National Family Health Survey, the wealth index is based on certain characteristics of housing and characteristics of a consumer items and the number is very large.

If you develop a single score on the basis of a large number of characteristics then what is done, that principal component analysis method or factor analytic method is used to arrive at weights which can be attached to different variables which emerge in the principal component analysis; this is what National Family Health Survey has done. So, for indexing, wealth index for building wealth index and for classification of people into different quantiles on the basis of wealth. This is just one example for many other purposes, these models or indexes can be used.

Estimation of what is not yet observed, quality and adjustment of data, we use when we were students, we used models; especially for quality and adjustment of age data from census age heaping, underestimation, overestimation in different age groups 0 to 4, 5 to 9, 0 to 9 and so on. Then studying causes and effects regression models. Nested models mean, that when you study relationship between say y and x 1, you want to be sure that the relationship is not due to x 2 x 3 x 4 and other variables; so nested models can be used and for modelling a phenomenal representation, and predictions.

(Refer Slide Time: 03:49)



Estimation of fertility (when data not available) – Coale's approach

$$TFR = \frac{P_3^2}{P_2}$$

If the fertility pattern can be described by a Gompertz function of the proportion experienced by each age, then

$$P_2 \left(\frac{P_4}{P_3}\right)^4$$

Bongaart's model to estimate fertility is discussed in another lecture

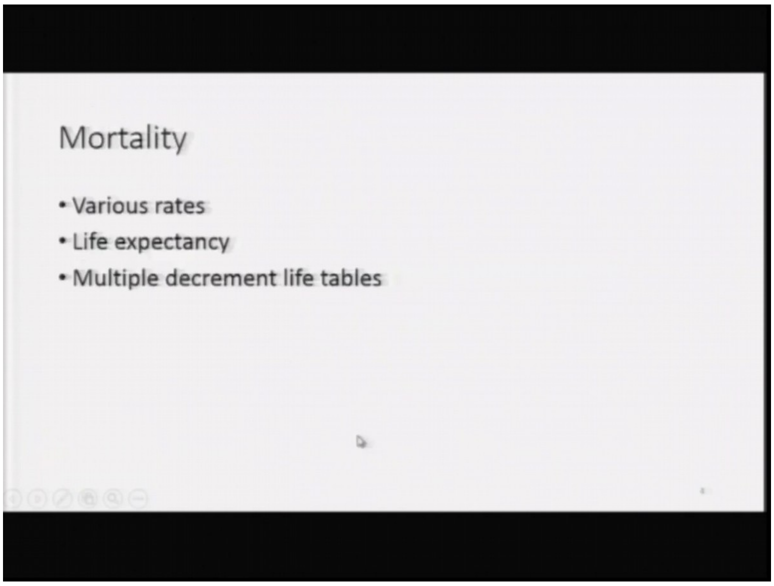
One example of modeling which was commonly used in population studies at one time, not now; because now you have direct estimates of total fertility rate from SRS and also from National Family Health Survey.

But when these data were not available, when SRS did not exist or when NFHS was not conducted; one of the methods used for estimation of total fertility rate was the TFR equal to P 3 square divided by P 2. What is P 3? P 3 is parity in the third reproductive age group, from 15 to 50; we divide the age period into 5 year age groups, 15 to 20, 20 to 25 to 25 to 29 or 30 depending on whether you define age as a discrete variable or continuous variable.

So, P 3 is parity; means average number of children born to women in the third 5 year reproductive age period; and P 2 is the parity number of children ever produced by women in the second age group in the reproductive period. If the fertility pattern can be described by Gompertz function of the proportion experienced by each age, then TFR was defined as P 2 into P 4 by P 3 raise to power 4.

Bongaart's model to estimate fertility which is more commonly used now, it discuss in another lecture; when I was talking about sociology of fertility and Kingsley Davis and Judith Blake's theory of 11 intermediate variables and Bongaart's model of estimation of fertility.

(Refer Slide Time: 05:45)

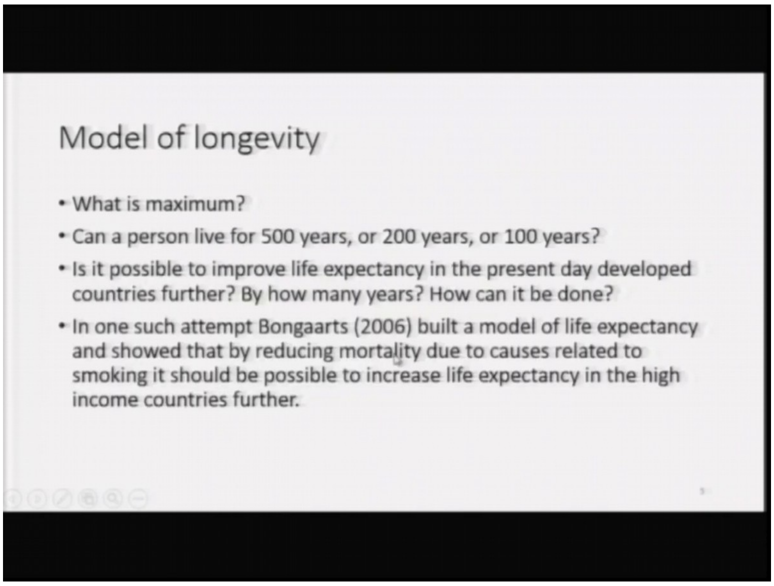


Mortality

- Various rates
- Life expectancy
- Multiple decrement life tables

Models are also used for mortality and I gave the example of life table, which is a stationary population model. It is used for computing various rates, life expectancy, and for making multiple decrement life tables.

(Refer Slide Time: 06:03)



Model of longevity

- What is maximum?
- Can a person live for 500 years, or 200 years, or 100 years?
- Is it possible to improve life expectancy in the present day developed countries further? By how many years? How can it be done?
- In one such attempt Bongaarts (2006) built a model of life expectancy and showed that by reducing mortality due to causes related to smoking it should be possible to increase life expectancy in the high income countries further.

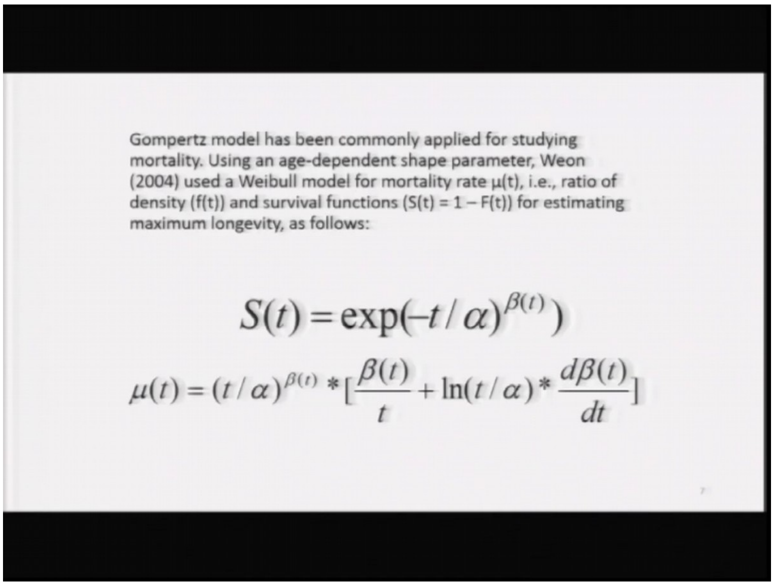
Model of longevity, you can ask questions like what is the maximum possible age? Can a person live for 500 years, or 200 years, or 100 years? We know that average life expectancy in some countries has risen to eighties, 82 years for females in Japan, for example. Is it possible to improve life expectancy in such countries further? Can average life expectancy be 85, 90, 95, 100; can it go beyond 100? This is what it means; can a person live for 500 years?

In one such attempt, Bongaart and this paper was published in population and development review built a model of life expectancy and showed that by reducing mortality due to causes related to smoking it should be possible to increase life expectancy in the high income countries further.

By decomposing, causes of mortality into various categories and by assuming at what ages further improvements are possible, by what interventions; Bongaart they found that, yes it is possible to increase life expectancy further and this article was written in 2006. I think somebody has to revise its exercise in 2019 now, and estimate weather Japan’s life expectancy can be raised to 90; and how?

So, that time by using a model which I will not go into detail, I will simply show what kind of equations or model they use. They showed that, yes it is possible to make some marginal improvements in life expectancy and that is possible because a number of deaths are associated with smoke and smoking practices if they are reduced, reduction in smoking practices can increase the life expectancy in developed countries also.

(Refer Slide Time: 08:31)



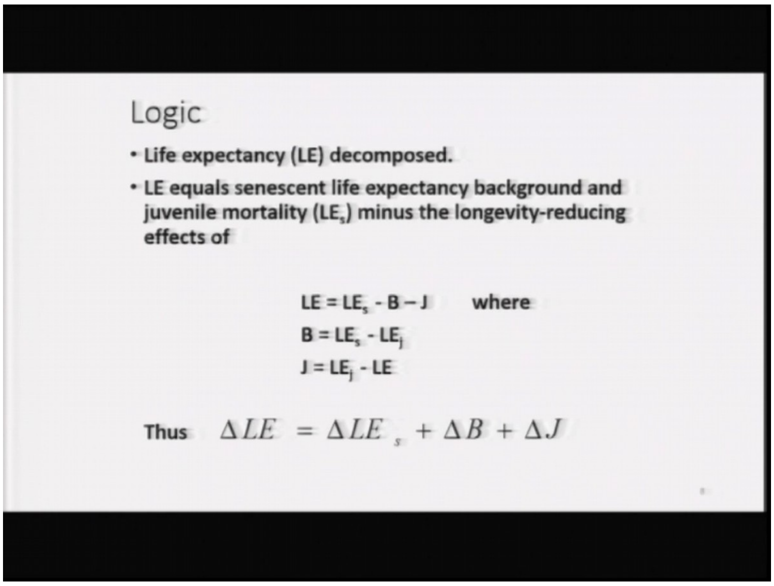
Gompertz model has been commonly applied for studying mortality. Using an age-dependent shape parameter, Weon (2004) used a Weibull model for mortality rate μ(t), i.e., ratio of density (f(t)) and survival functions (S(t) = 1 – F(t)) for estimating maximum longevity, as follows:

$$S(t) = \exp(-t/\alpha)^{\beta(t)})$$

$$\mu(t) = (t/\alpha)^{\beta(t)} * [\frac{\beta(t)}{t} + \ln(t/\alpha) * \frac{d\beta(t)}{dt}]$$

They use these models, Gompertz model.

(Refer Slide Time: 08:37)



Logic

- **Life expectancy (LE) decomposed.**
- **LE equals senescent life expectancy background and juvenile mortality (LE_s) minus the longevity-reducing effects of**

$LE = LE_s - B - J$ **where**

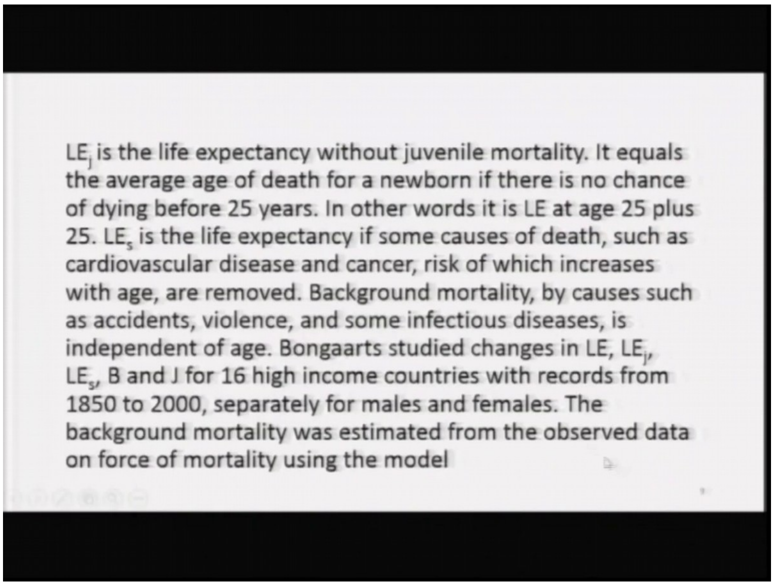
$B = LE_s - LE_j$

$J = LE_j - LE$

Thus $\Delta LE = \Delta LE_s + \Delta B + \Delta J$

And they also require some logic like life expectancy LE decomposed. LE equals senescent life expectancy background and juvenile mortality LEs minus the longevity reducing effects of LE equal to LE s minus B minus J; B is LE s minus LE j; J equal to LE j minus LE. Thus change in LE equal to change in LE s plus change in B plus change in J and by using this equation, then they found out how much change in life expectancy is possible.

(Refer Slide Time: 09:21)

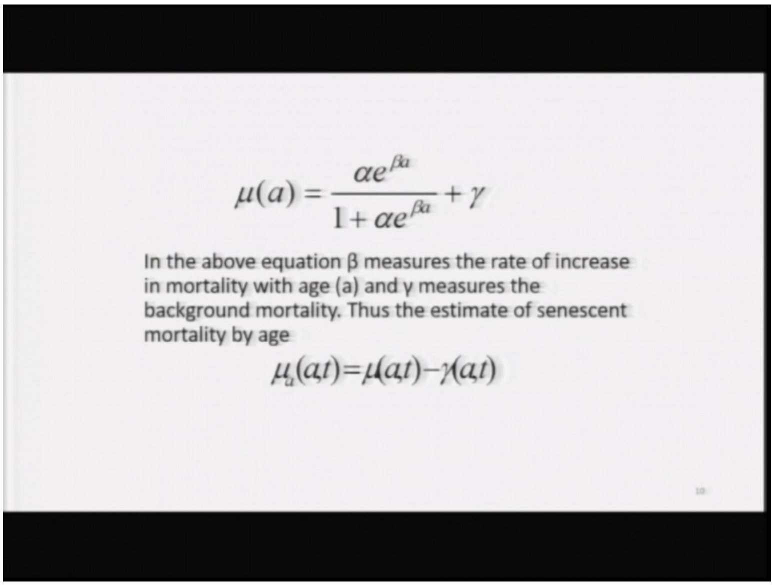


LE_j is the life expectancy without juvenile mortality. It equals the average age of death for a newborn if there is no chance of dying before 25 years. In other words it is LE at age 25 plus 25. LE_s is the life expectancy if some causes of death, such as cardiovascular disease and cancer, risk of which increases with age, are removed. Background mortality, by causes such as accidents, violence, and some infectious diseases, is independent of age. Bongaarts studied changes in LE, LE_j, LE_s, B and J for 16 high income countries with records from 1850 to 2000, separately for males and females. The background mortality was estimated from the observed data on force of mortality using the model

LE j is the life expectancy without juvenile mortality. It equals the average age of death for a newborn baby if there is no chance of dying before 25 years. In other words it is LE at age 25 plus, LE s is the life expectancy if some causes of death, such as cardiovascular disease and cancer, risk of which increases with age, are removed. If these diseases or these causes of death are removed, what will happen to life expectancy? Background mortality by causes such as accidents, violence, and some infectious diseases is independent of age.

So, they made the assumption that certain causes of death depend on age and certain others are independent of age. Bongaart's studied changes in LE, LE j, LE s, B and J for 16 high income countries with records from 1850 to 2000 a long period of time 150 years, separately for males and females. The background mortality was estimated from the observed data on force of mortality using the model.

(Refer Slide Time: 10:39)



$$\mu(a) = \frac{\alpha e^{\beta a}}{1+\alpha e^{\beta a}} + \gamma$$

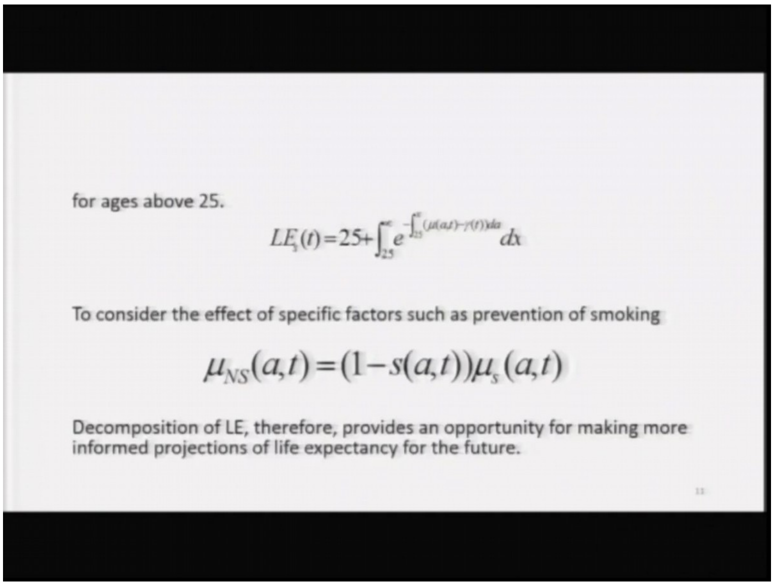
In the above equation β measures the rate of increase in mortality with age (a) and γ measures the background mortality. Thus the estimate of senescent mortality by age

$$\mu_s(a,t)=\mu(a,t)-\gamma(a,t)$$

10

Mu a equal to alpha e raised to power beta a divided by 1 plus alpha e raised to power beta a plus gamma. In the above equation, beta measures the rate of increase in mortality with age a, and gamma measures the background mortality. Thus the estimate of senescent mortality by age is this, mu a a t equal to mu a t minus gamma a t. It can be quite confusing for those who do not have the right background in mathematics, so I am avoiding this.

(Refer Slide Time: 11:09)



for ages above 25.

$$LE_s(t)=25+\int_{25}^{\infty} e^{-\int_{25}^{x}(\mu(a,t)-\gamma(t))da}dx$$

To consider the effect of specific factors such as prevention of smoking

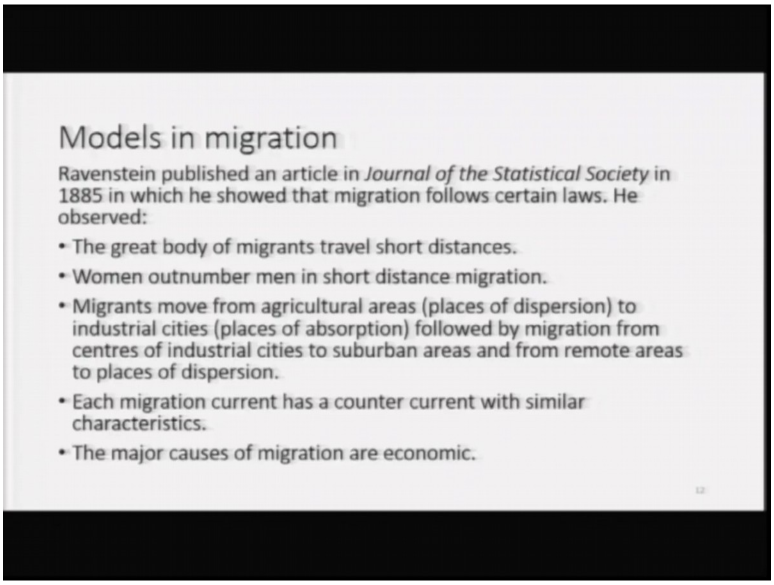
$$\mu_{NS}(a,t)=(1-s(a,t))\mu_s(a,t)$$

Decomposition of LE, therefore, provides an opportunity for making more informed projections of life expectancy for the future.

11

I just wanted to show, that an interesting question like by how many more years life expectancy can be increased in developed countries, can be answered by dividing deaths into age independent, age dependent factors, and using certain models in terms of force of mortality.

(Refer Slide Time: 11:43)



Models in migration

Ravenstein published an article in *Journal of the Statistical Society* in 1885 in which he showed that migration follows certain laws. He observed:

- The great body of migrants travel short distances.
- Women outnumber men in short distance migration.
- Migrants move from agricultural areas (places of dispersion) to industrial cities (places of absorption) followed by migration from centres of industrial cities to suburban areas and from remote areas to places of dispersion.
- Each migration current has a counter current with similar characteristics.
- The major causes of migration are economic.

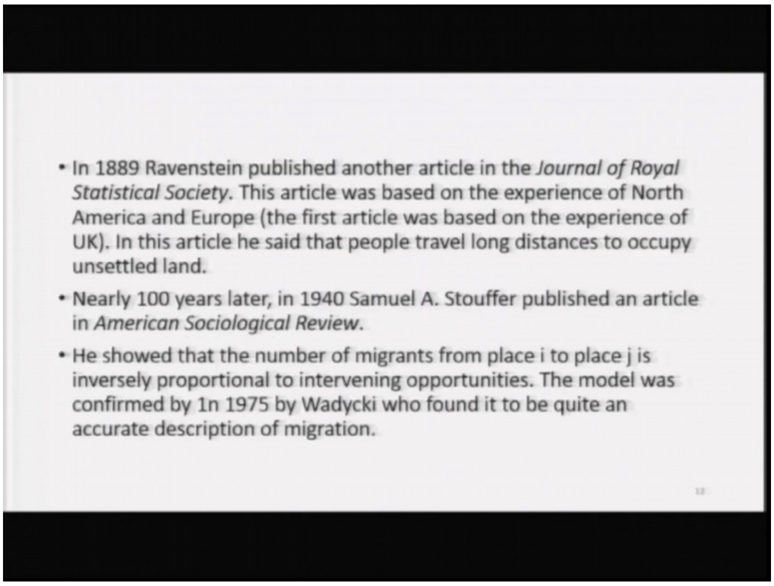
12

Models have been used in migration also. So, fertility, proximate determinants of fertility is one example of modeling; in mortality this Bongaart's model this is an example of modelling in mortality.

Now, models in migration, migration also makes use of modeling. Ravenstein published an article in Journal of Statistical Society in 1885 and still Ravenstein model or Ravenstein model of migration is still referred to literature on migration. In which he showed that migration follow certain laws, he observed: the great body of migrants travel short distances, women outnumber men in short distance migration, migrants move from agricultural areas to industrial cities of places of absorption, rural areas are places of dispersion, urban areas or industrial cities are places of absorption followed by migration from centers of industrial cities to suburban areas.

This is the next stage of migration. Migration takes place first from rural areas to urban area; and then subsequently from a heart of the city, core of the city to peripheral or suburban areas and from remote areas to places of dispersion. Each migration current has a counter current with similar characteristics; the major causes of migration are economy.

(Refer Slide Time: 13:23)



- In 1889 Ravenstein published another article in the *Journal of Royal Statistical Society*. This article was based on the experience of North America and Europe (the first article was based on the experience of UK). In this article he said that people travel long distances to occupy unsettled land.
- Nearly 100 years later, in 1940 Samuel A. Stouffer published an article in *American Sociological Review*.
- He showed that the number of migrants from place i to place j is inversely proportional to intervening opportunities. The model was confirmed by 1n 1975 by Wadycki who found it to be quite an accurate description of migration.

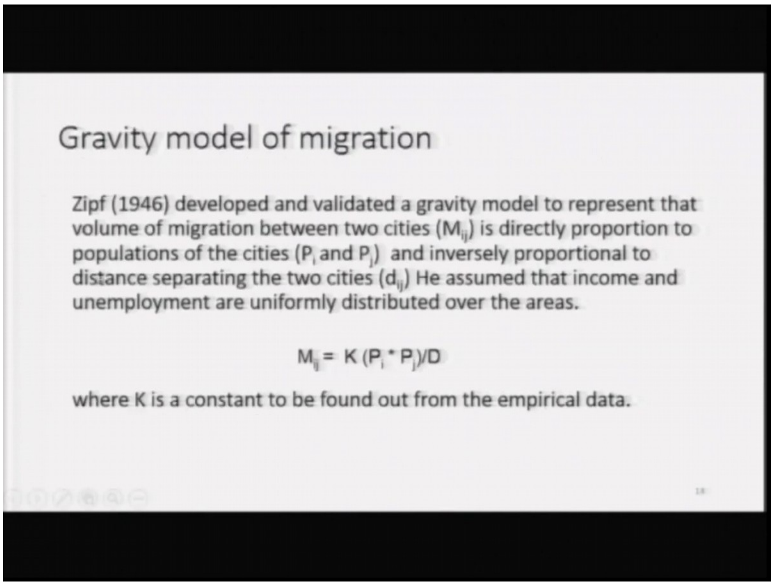
12

In 1889 Ravenstein published another article in the Journal of Royal Statistical Society. This article was based on the experience of North America and Europe, the first article was based on the experience of UK.

In this article he said that people travel long distances to occupy unsettled land. Nearly 100 years later, in 1940 Samuel A Stouffer published an article in American Sociological Review. He showed that the number of migrants from place i to place j is inversely proportional to intervening opportunity. In one lecture I was referring to Zipf’s model of migration, according to which migration equal to population of place of origin multiplied by population of place of destination divided by distance separating the two and there is a constant of proportionality.

Now, this model, Stouffers model also talks of intervening opportunities; if there are more intervening opportunities, then people need not go too far off places. The model was confirmed in 1975 in more accurate descriptions of migration, means it is still valid.

(Refer Slide Time: 14:45)



Gravity model of migration

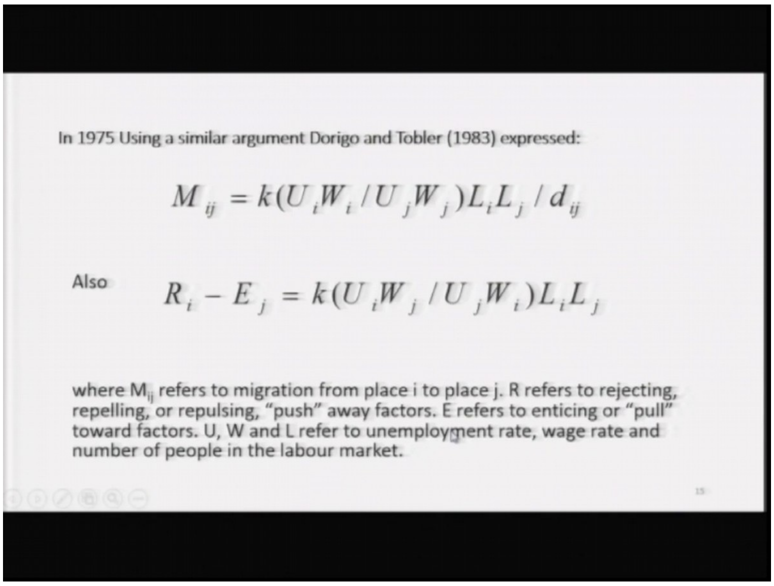
Zipf (1946) developed and validated a gravity model to represent that volume of migration between two cities (M_{ij}) is directly proportion to populations of the cities (P_i and P_j) and inversely proportional to distance separating the two cities (d_{ij}) He assumed that income and unemployment are uniformly distributed over the areas.

$$M_{ij} = K (P_i * P_j)/D$$

where K is a constant to be found out from the empirical data.

The Zipf's model is also called gravity model of migration and this is what I was talking about.

(Refer Slide Time: 14:51)



In 1975 Using a similar argument Dorigo and Tobler (1983) expressed:

$$M_{ij} = k(U_i W_i / U_j W_j) L_i L_j / d_{ij}$$

Also $$R_i - E_j = k(U_i W_j / U_j W_i) L_i L_j$$

where M_{ij} refers to migration from place i to place j. R refers to rejecting, repelling, or repulsing, "push" away factors. E refers to enticing or "pull" toward factors. U, W and L refer to unemployment rate, wage rate and number of people in the labour market.

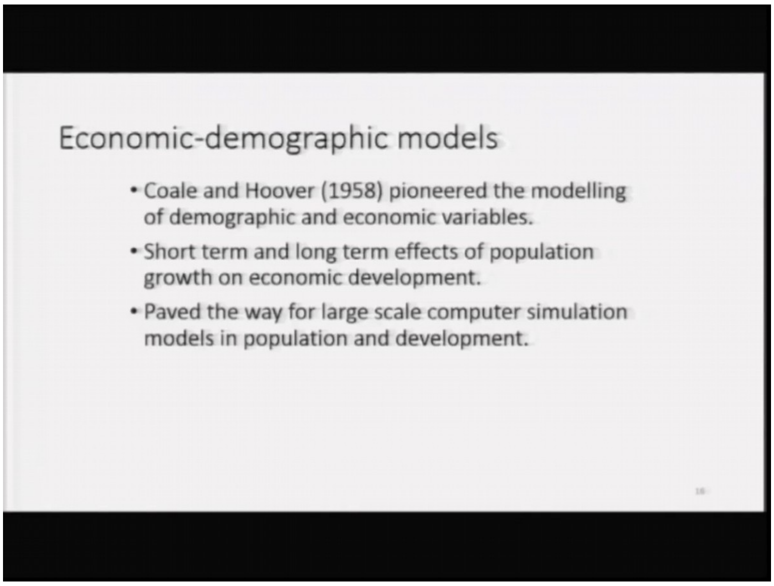
Now, in 1975 using a similar argument Dorigo and Tobler expressed M i j, M i j is migration from place i to j, that is constant of proportionality k. U i is unemployment rate at i th place, W i is wage rate at i th place divided by U j unemployment rate at j th place, W j wage rate at j th place, L i is size of labour force at i th place, and L j is size of labour force at the j th place and this thing divided by d i j.

So, what has happened that in place of population, they have used labour force, size of labour force. If you forget this part U i W i by U j W j, then it is L i L j by d this is same as Zipf; except that population is replaced by size of labour force. And this part, unemployment rate and wage rate they determine whether a place of destination as compared to place of origin will be attractive.

So, they also talk of repelling forces, and push away factors, and enticing or pull factors, and the difference between R and E means rejecting factors, and enticing factors is defined as k U i W j divided by U j W i into L i L j.

Whether this model applies to all kinds of migration streams or not; that is not the point here. I just wanted to show that in the studies of migration also modelling has been used.

(Refer Slide Time: 16:53)



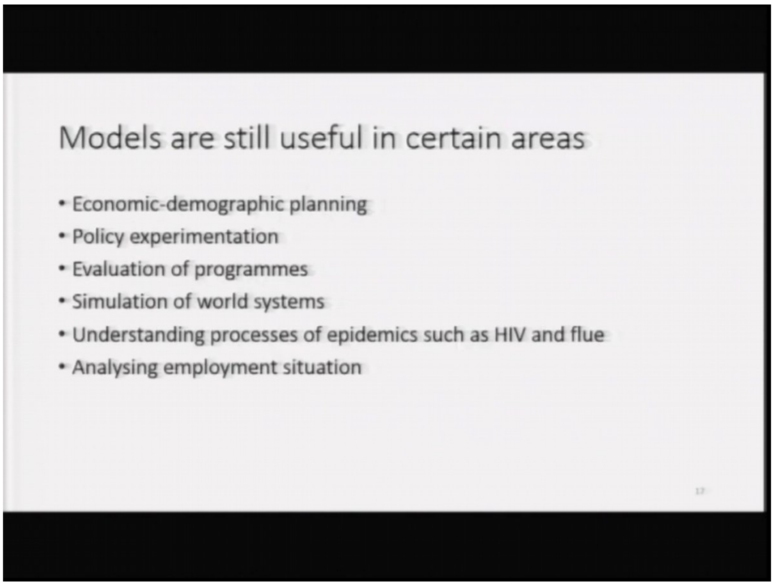
Economic-demographic models

- Coale and Hoover (1958) pioneered the modelling of demographic and economic variables.
- Short term and long term effects of population growth on economic development.
- Paved the way for large scale computer simulation models in population and development.

16

Then there are economic demographic models; Coale and Hoover in 1958 pioneered the modelling of demographic and economic variables. They studied short term and long term effects of population growth on economic development. And they paved the way for large scale computer simulation models in population and development which became more popular subsequently.

(Refer Slide Time: 17:17)



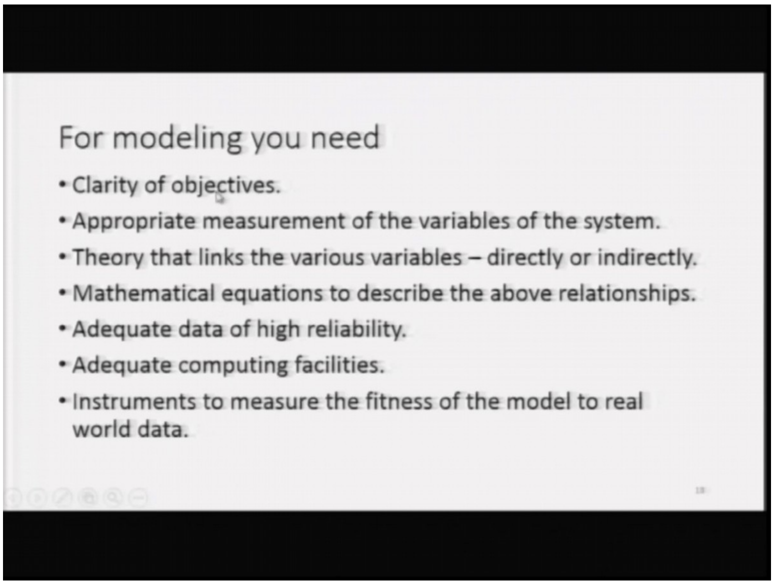
Models are still useful in certain areas

- Economic-demographic planning
- Policy experimentation
- Evaluation of programmes
- Simulation of world systems
- Understanding processes of epidemics such as HIV and flue
- Analysing employment situation

17

Models are still used in certain areas; economic demographic planning, policy experimentation, evaluation of programs, simulation of word systems, understanding processes of epidemics such as HIV and flue. Many mathematics department, and many mathematicians specializing in bio demography are using such models to study epidemic such as HIV and analyzing employment situation.

(Refer Slide Time: 17:55)



For modeling you need

- Clarity of objectives.
- Appropriate measurement of the variables of the system.
- Theory that links the various variables – directly or indirectly.
- Mathematical equations to describe the above relationships.
- Adequate data of high reliability.
- Adequate computing facilities.
- Instruments to measure the fitness of the model to real world data.

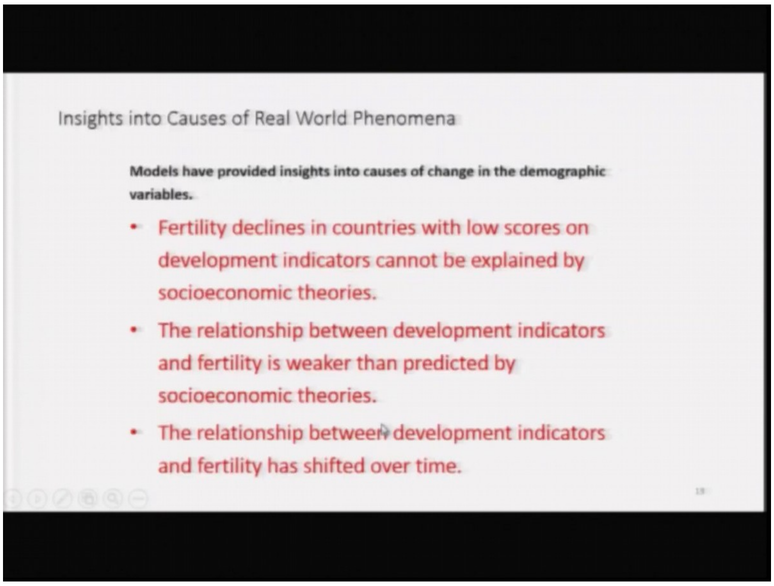
18

For modeling, what you need is clarity of objective; because models are simple models are complex and depending on the purpose you develop different types of models.

So, clarity of objectives; what is the objective of modeling? Appropriate measurement of the variables of the system; means appropriate measurement of the variables which are included in the model. Then theory, you need a theory that links the various variables directly or indirectly, this understanding of theory will come from review of literature.

Then you need mathematical equations to describe the above relationships, then you need adequate data of high reliability, adequate computing facilities and instruments to measure the fitness of the model to real world data.

(Refer Slide Time: 18:51)



Insights into Causes of Real World Phenomena

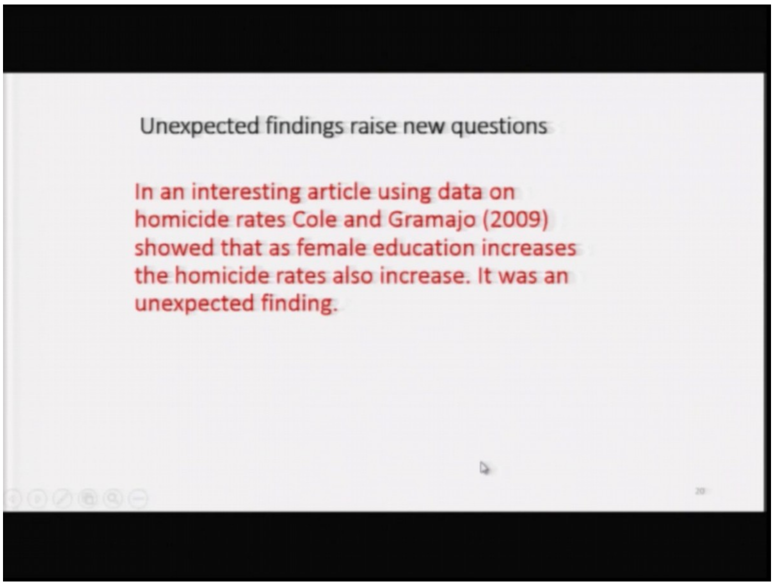
Models have provided insights into causes of change in the demographic variables.

- Fertility declines in countries with low scores on development indicators cannot be explained by socioeconomic theories.
- The relationship between development indicators and fertility is weaker than predicted by socioeconomic theories.
- The relationship between development indicators and fertility has shifted over time.

This insights into causes of real world phenomenon require, that models have you know models have provided insights into causes of change in the demographic variables.

And here are some examples of what models have said; fertility declines in countries with low scores on development indicators cannot be explained by socioeconomic theories. Modelling has shown that the relationship between development indicators and fertility is weaker than predicted by socioeconomic theories and the relationship between development indicators and fertility has shifted overtime.

(Refer Slide Time: 19:37)



Unexpected findings raise new questions

In an interesting article using data on homicide rates Cole and Gramajo (2009) showed that as female education increases the homicide rates also increase. It was an unexpected finding.

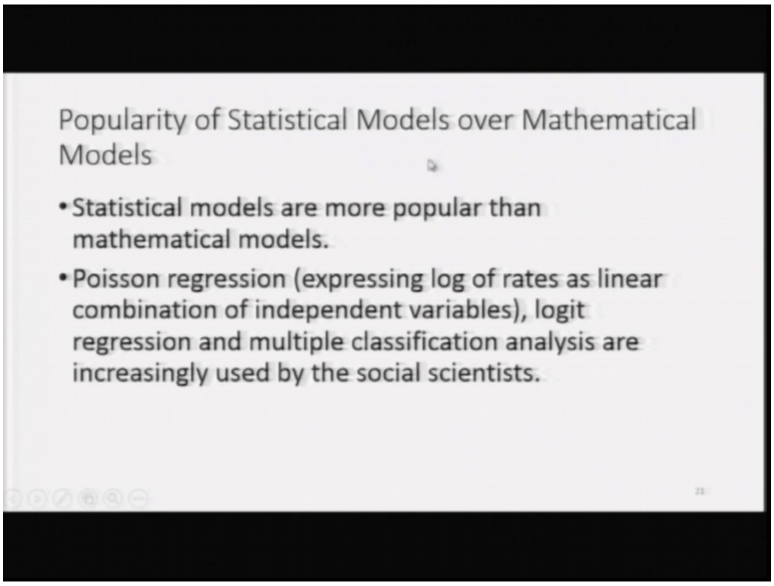
Unexpected findings raise new questions; in an interesting article using data on homicide rates Cole and Gramajo. In 2009 this article was also published in population and development review, showed that as female education increases the homicide rates also increase. It was an unexpected finding.

I often discuss this article in my sociology courses and generally sociology courses start with, studies of suicide by Emile Durkheim. I say that, if in that age of Emile Durkheim, only simple algebra was known, simple mathematics was known ratios rates. If Emile Durkheim would have been alive, then he could have done what Cole and Gramajo did in 2009.

They collected data on homicide rates for different countries, identified some variables like literacy among males, literacy among females, some variables of culture, some variables of industrialization and urbanization, some variables of urbanization and so on, and use the method of logistic regression.

By using logistic regression they found that, not male education; but female education determines the homicide rates in a country; and then they gave a number of possible reasons, why female education can explain homicide rates.

(Refer Slide Time: 21:21)



Popularity of Statistical Models over Mathematical Models

- Statistical models are more popular than mathematical models.
- Poisson regression (expressing log of rates as linear combination of independent variables), logit regression and multiple classification analysis are increasingly used by the social scientists.

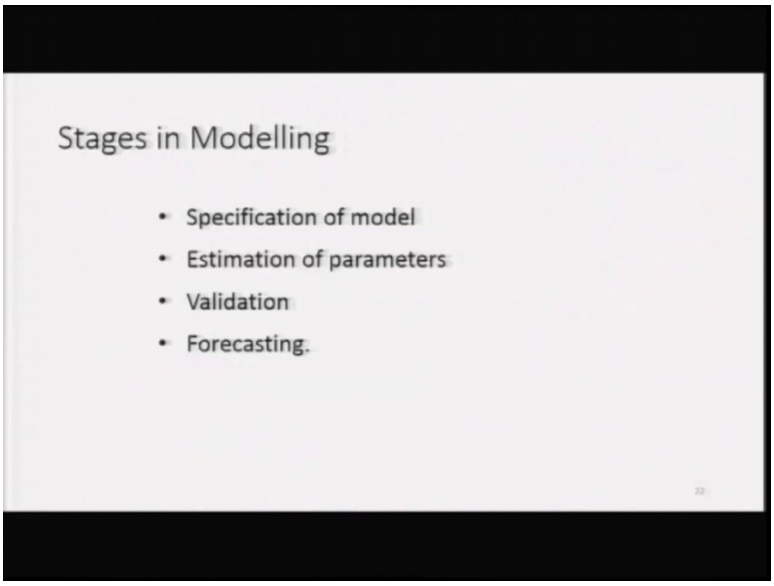
21

Statistical models are certainly more popular than mathematical model, because like a poisson regression; expressing log of rates a linear combination of independent variables, logit regression and multiple classification analysis are increasingly used by the social scientists. The multiple classification analysis was used to analyze fertility differences more in 1980 and 90.

Nowadays I have not seen use of multiple classification energy; these days we are going more for other techniques such as path analysis, logistic regression, and structural equation model.

And in regression analysis, one term I mentioned poisson regression. Poisson regression is another type of regression analysis, something akin to logistic regression. Under certain conditions of dependent and independent variables we use different types of logit, prohibit, poisson, or logistic regression.

(Refer Slide Time: 22:41)

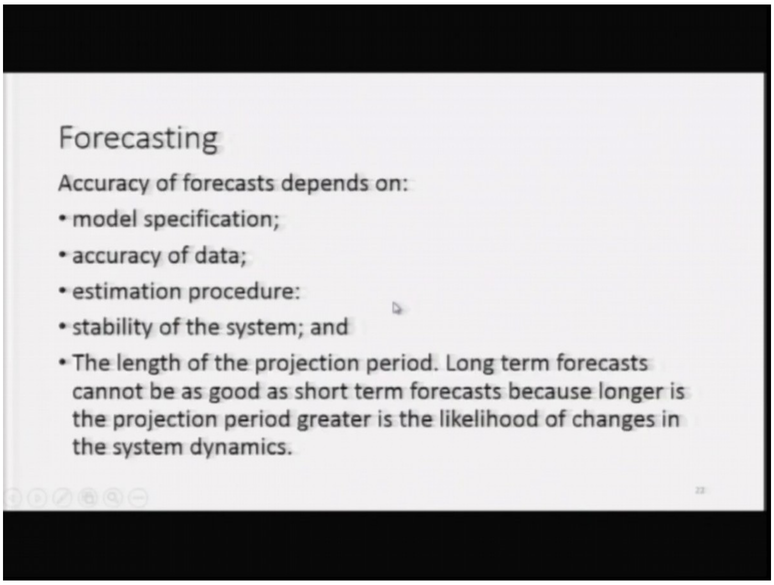


Stages in Modelling

- Specification of model
- Estimation of parameters
- Validation
- Forecasting.

There are certain stages in modeling; first is specify the model, then estimate it is parameters, then validate it, and then forecast. Validation means, answering to what extent your model is ok.

(Refer Slide Time: 22:53)



Forecasting

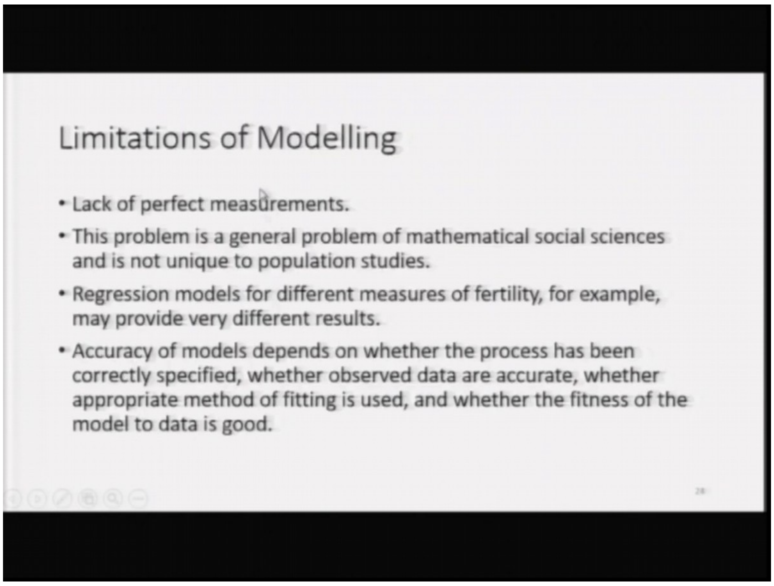
Accuracy of forecasts depends on:

- model specification;
- accuracy of data;
- estimation procedure:
- stability of the system; and
- The length of the projection period. Long term forecasts cannot be as good as short term forecasts because longer is the projection period greater is the likelihood of changes in the system dynamics.

Forecasting accuracy depends on: model specification; accuracy of data; estimation procedure; stability of a system; and the length of the projection period. It does not depend on what kind of model you are building alone; it depends on a variety of factors.

Long term forecast cannot be as good as short term forecast because the patterns of relationships might change, because longer is the projection period greater is the likelihood of changes in the system dynamics.

(Refer Slide Time: 23:27)



Limitations of Modelling

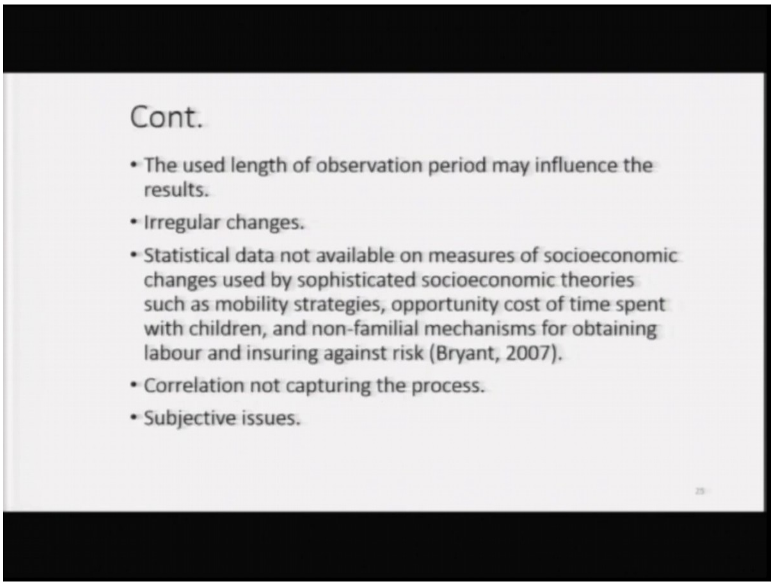
- Lack of perfect measurements.
- This problem is a general problem of mathematical social sciences and is not unique to population studies.
- Regression models for different measures of fertility, for example, may provide very different results.
- Accuracy of models depends on whether the process has been correctly specified, whether observed data are accurate, whether appropriate method of fitting is used, and whether the fitness of the model to data is good.

There are some limitations of the modeling; lack of perfect measurements. This problem is a general problem mathematical social sciences and is not unique to population studies.

Regression models for different measures of fertility, for example, may provide very different results you may measure fertility by crude birth rate, you may measure it by total fertility rate, gross reproduction rate, you may measure it by net reproduction rate, as dependent variable; and the results will be different.

Accuracy of models depends on whether the process has been correctly specified, whether observed data are accurate, whether appropriate method of fitting is used, and whether the fitness of the model to data is good; for that again there are so many statistics.

(Refer Slide Time: 24:15)



Cont.

- The used length of observation period may influence the results.
- Irregular changes.
- Statistical data not available on measures of socioeconomic changes used by sophisticated socioeconomic theories such as mobility strategies, opportunity cost of time spent with children, and non-familial mechanisms for obtaining labour and insuring against risk (Bryant, 2007).
- Correlation not capturing the process.
- Subjective issues.

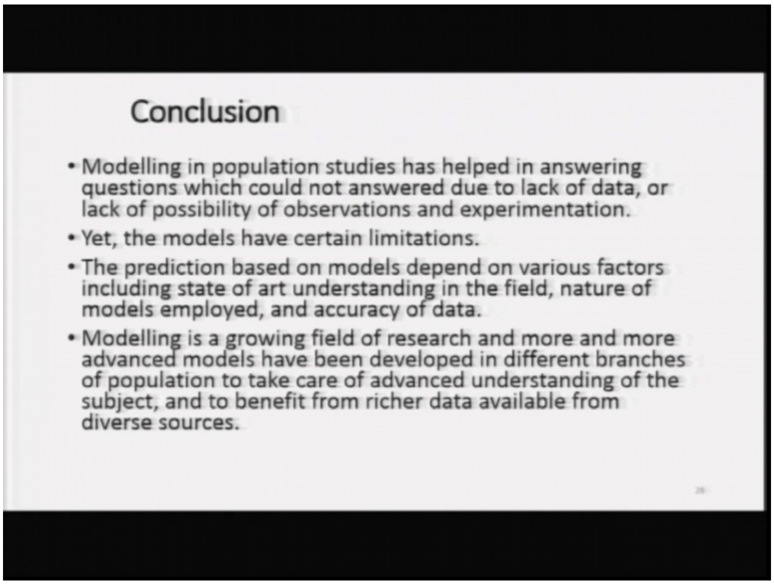
25

The used length of observation period may influence the results.

And there may be regular changes, irregular changes; statistical data are not available on measures of socio economic changes used by sophisticated socio economic theory such as mobility strategies, opportunity cost of time spent with children, and non familial mechanisms for obtaining labour and insuring against risk.

Correlation does not capture the process and there are also subjective issues, which have not been properly measured.

(Refer Slide Time: 24:55)



Conclusion

- Modelling in population studies has helped in answering questions which could not answered due to lack of data, or lack of possibility of observations and experimentation.
- Yet, the models have certain limitations.
- The prediction based on models depend on various factors including state of art understanding in the field, nature of models employed, and accuracy of data.
- Modelling is a growing field of research and more and more advanced models have been developed in different branches of population to take care of advanced understanding of the subject, and to benefit from richer data available from diverse sources.

26

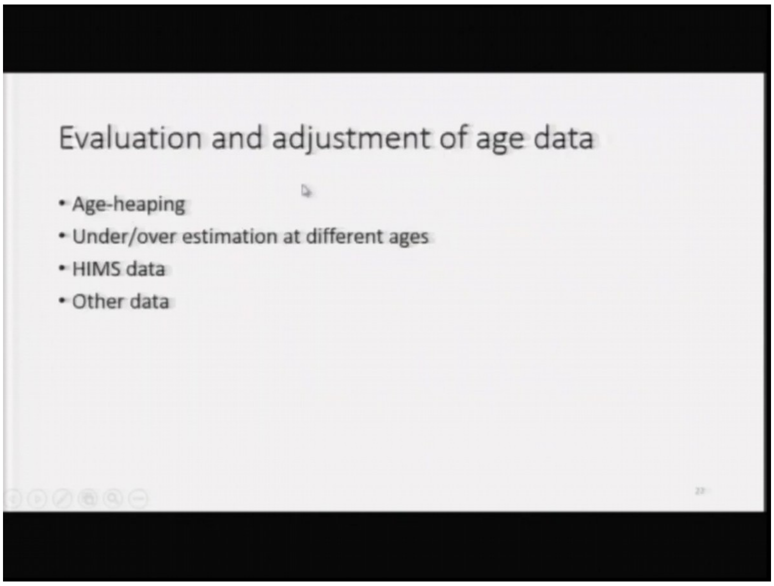
Conclusion, modelling in population studies has helped in answering questions which could not be answered due to lack of data. How many people have ever lived on earth cannot be answered; but if you know that during different periods of time there have been different rates of growth of population and the population growth followed a particular model.

It may be a simple model like exponential, then for different periods you can calculate how many people have lived; and by adding people lived in different periods you arrive at the figure and I mention that the figure would be closed to 100. So, 7.7 percent of the people who have ever lived on this planet earth, we are or we are living; and the models have certain limitations the prediction based on models depend on various factors.

I remember that when I was a student in 77 or 78, we wrote a paper in Economic Times of Bombay. Bombay edition of economic times; that what is the reason why fertility was not declining at that time; all the people wanted 2 sons and 1 daughter; and we said that 2 sons and 1 daughter will not add to 3, it will add to 5 because all couples will not be so fortunate as to have 2 sons and 1 daughter out of 3, they will have to produce more.

And then due to replacement hypothesis, under the condition of high mortality it makes sense to have 5. Modelling is a growing field of research and more and more advanced models have been developed in different branches of population to take care of advanced understanding of the subject, and to benefit from richer data available from diverse source.

(Refer Slide Time: 26:55)



Evaluation and adjustment of age data

- Age-heaping
- Under/over estimation at different ages
- HIMS data
- Other data

Modelling is also used in evaluation and adjustment of age data, for example, age-heaping I mentioned about this; under and over estimation at different ages and HIMS data and other data.

Thank you.