Seminar of Andrei @ UTD: first session

Meisam Hejazinia

1/14/2013

optional text books:

Gelman, carlin, stern, and rubin (2003), Bayesian data analysis (2nd eddition), chapman and Hall
Greene (2006), Econometric analysis (65h edition), prentice hall

midter 30%
term paper 60%
term paper presentation 10%

Some analysis should be done in the paper, and some bayesian estimation in it.

You need to provide the written description of the data next week. You want to dirty your hand with data.

Last class would be presentation.

Data should be available in term of verifying the data.

Introduction:

Why do we need statistical analysis?

DEA vs. Stochastic Frontier

Uncertainty

Common Approaches to statistical analysis

Frequentiest methods:

- Least Squares
- Maximum Likelihood
- Example of derivations for a linear regression

Bayesian Approach :

- Explicit use of probability for quantifying uncertainty in inferences
- Analogy with driverless car competition

Model performance is a measure for quality of the model.

Possible explanation for R-square is the sporious relation, or not capturing the essential process behind it.

The car race, which was about frequentist approach, and as they changed their approach by assuming the distribution for them then they could finish.

Bayesian Data Analysis:

set up a full probability model:

Joint probability distribution for all observable and unobservable quantities in a problem.

model should be consistent with the knowledge about the underlying problem and data collection.

Conditional on observed data, calculate the appropriate posterior distribution of the unobserved

quantities of ultimate interest

Evaluate the fit and interpret the implications of the resulting posterior distribution:

Does the model fit the data

Are the substantive conclusiosn reasonable

How sensitive are the result to modeling assumptions in step 1

Usually frequentists question the assumption of having the prior

When you have a dice you can come with the prior, and then the question will come that whether it is fair or not. As a result it is not always difficult to know the prior.

Markov chain is used for this purpose that you plug in the priors in.

Markove chain allows you to separate the probabilities and get estimate simpler.

You should always be worried about whether your estimate is reasonable.

What does confidence interval means in frequentist approach?

Common sense interpretations of statisitcal conclusions:

Bayesian (probability) interval for an unknown quantity of interest

Directly implies a high probability of containing the unknown quantity

Vs. Frequentist (confidence) interval

Implies a sequency of similar inferences made in repeated practices

Logit can be approximated by probit.

If you don't care about six degree of freedom you can use probit, and you will get the logit estimation.

You should be able to justify the method you use.

General Notions:

In general, statisitcal inference attemps to draw condlusion about quantities that are not observed (estimands) based on numerical data. Estimands come in two favors: a) potentially observable (e.g. future observations), and b) not directly observable (e.g. regression coefficients):

θ population parameters (vector of unboservable quantities)
y : observed data
$\bar{y}$ potentially observable quantities
α, β, γ parameters
x, w observable/ observed scalars and vectors
X, W observable/ observed matrices

When you can see and count something you have observed that thing.

You assume some to be known, and you draw unknown things out of those observables.

Example of unobserved is the coefficient of gravity, but you can count the drops, so it is observable.

Random variables:

Outcome variables y are considered random: Each observed value could have turned out different due to the sampling process and natural variation in population

$y = (y_1, y_2, \ldots, y_n)$

Joint probability density p(y) is invariant to permutations of the indexes (exchangeability)

Usually data from an exchangeable distribution is modled as i.i.d given some unknown parameter vector θ

Bayesian Inference

Conclusions about parameters are made in terms of probability statements:

$p(\theta|y)$

This probability is explicitly conditional on observed outcomes, and implicitly conditional on any covariate

$p(\theta, y) = p(\theta)p(y|\theta)$

$p(\theta, y) = \frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$

$p(y) = \sum p(\theta)p(y|\theta)$

Since everything would sum up to one, then $p(\theta|y) \propto p(\theta)p(y|\theta)$

Bayesian Inference:

$p(\theta|y) \propto p(\theta)p(y|\theta)$

This is the technical essence of Bayesian inference:

1. develop the model $p(\theta, y)$, and
2. perform the necessary computation to summarize $p(\theta|y)$

You start with normal distribution, and any distribution and you simulate multiple random numbers.

Markove chain and random number generation has nothing to do with Bayesian, but you need to be able to do that on MATLAB.

Posterior odds:

For the next session provide the written description of data that you are interested in. You should try to work on the data that you like and understand to run your homeworks on.

The good book is "Greenberg" book of "introduction to Bayesian statistics".

Uniform prior put huge weight on the tails. Typically if you have the data of what the piror should be, try to use it as prior rather than uniform.

Homophilia example

Males have X-Y chromosomes, and females: X-X chromosomes

Homophilia is inhterited through the X-chromosome

Males are affected with just one "bad" chromosome, but females are affected only with two bad genes

Consider a woman who has affected brother and unaffected parents, which implies that her mother has one "bad" chromosome

The unknown quantity of interest is whether the women is herself a carrier of the "bad" gene $(\theta = 1), ornotcarrier(\theta = 0)$

The prior distribution based on presented information is:

$p(\theta = 1) = p(\theta = 0) = 0.5$
The modl and likelihood:

Outcome variable $y_t = 0$ if son i is not affected and $y_t = 1$ if son i is affected

The woman has two unaffected sons who are not identical twins:

$y_1 = 0, y_2 = 0$

$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$

$p(\theta = 1|y_1 = 0, y_2 = 0) =$
Transformation of variables:

Suppose p(u) is the density of vector u, and we need a transformation v=f(u)

Discrete distribution p(.):

Density $q(v) = p(f^{-1}(v))$

Continuous distribution p(.):

Density $q(v) = |J|p(f^{-1}(v))$

Where $|J|$ is the Jacobian of the transformation $u = f^{-1}(v)$

Example for univariate normal distribution:

If $u \sim N(\mu, \sigma^2)$ then what is distribution of $v = \frac{u-\mu}{\sigma}$

$J = \frac{\partial u}{\partial v} = \sigma$

Need review of Jacobian. This is the kind of stuff that will be asked on midterm.

$E(u) = \int\int up(u,v)dudv = \int\int u.p(u|v)dup(v)dv = \int E(u|v)p(v)dv = E(E(u|v))$

$E(\theta) = E(E(\theta|y))$

$var(\theta) = E(var(\theta|y)) + var(E(\theta|y))$

Difference between frequentist approach and Bayesian approach:

1. Diffference in philosophy

2. Ease of calculation

It is not some magical test, but it is only 1. different philosophy and 2. ease of calculation.

Both these approaches sound eminently reasonable, to the point that differences between them sound subtle to the point of unimportance.

Frequentist: The world is a certain way, but I dont know how it is. Further, I cant necessarily tell how the world is just by collecting data, because data are always finite and noisy. So Ill use statistics to line up the alternative possibilities, and see which ones the data more or less rule out., they calculate $P(D|H)$, treating data as random, and hypothesis as fixed, and H is some "null" hypothesis.
A frequentist interprets the word probability as meaning the frequency with which something would happen, in a lengthy series of trials
Frequentist say that data is consistent with the probability of life in mars.

A Bayesian basically says, I dont know how the world is. All I have to go on is finite data. So Ill use statistics to infer something from those data about how probable different possible states of the world are. they calculate, $P(H|D)$.
Bayesian interpretation of probability (though not the only one) is as subjective degree of belief: the probability that you (personally) attach to a hypothesis is a measure of how strongly you (personally) believe that hypothesis.
Bayesian would say Theres probably not life on Mars

There are contexts in which Bayesian and frequentist statistics easily coexist.

Seminar of Andrei @ UTD: Second session

Meisam Hejazinia

1/28/2013

The grading would be on methodology of the project and not the question. The week after the spring break would be first presentation of the project.

Observed heterogeneity could be captured from the individual characteristics, or through covariates

DEA: Data Envelop Analysis: helps you check the efficiency frontier

Two effects of random effect or fixed effect. Fixed effect, you would think there would be intercept. It is important to not capture fixed, but a random effect as well.

Bayesian can merge them and do that simply, assuming that each have different variable different for individual. It would not be complicated, but be more simple.

Heterogeneity in Basian would be random effect. In frequentist is that β_i would be different with a distribution. The data would be of the form of panel. You just will do simulation. You would not have the problem to check whether it is local maximum or global. It may take over night for calculation, but would be simple.

It is not important what method of estimation you use. Random number generation for simulation by computer based on the distribution. We use MCMC in bayesian and split it into separate pieces and use markov chain to see what we can take, and they will converge.

In frequentist contineous and discrete nature of the variabel would be different. Splitting means you condition on μ and σ^2 and simulate each of them.

In the bayesian after simulation you know the distribution of random variable.

Information prior distribution

Population interpretation: all possible parameter values

State of knowledge interpretation: use subjective knowledge about parameters

Prior should include all plausible parameter values, but need not be concentrated around true value

Conjugate prior distirbution

If F is class of sampling distribution $p(y|\theta)$ and P is class of prior distributions for θ, then the class P is conjugate for F if:

$P(\theta|y) \in P$ for all $p(.|\theta) \in F$ and $p(.) \in P$

Exponential family distribution

Only classes of distribution that have natural conjugate priors

$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)'u(y_i)}$

$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

$e^{\frac{1}{2\sigma^2}[-1-2\mu][-\frac{2\mu}{2\sigma^2}}$ $^{-\frac{1}{2\sigma^2}}$

$\phi = [-\frac{1}{2\sigma^2} - \frac{2\mu}{2\sigma^2}]$

Prior distribution with no population basis: Vague, flat, diffuse, non-informative

Proper prior density: Independent of data and integrates to one
Improper priors can lead to proper posterior

Conjugate could be taken when you are within the exponential family.

Posterior for a conjugate prior could be calculated by the form of $p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}B(y+\alpha, n-y+\beta)$. α and β are hyper parameters here.

Predicting in sample, and out of sample. The prior should be plausible.

t-distribution, fatter tail, and for lower sample size.

Parameter itself is nothing, and you need to look at the marginal effect.

$\frac{1}{\tau^2}$ is called perceision, which is reverse of variance.

Exponential family

Matrix algebra

n: number of observation for exponential model

$\bar{y}$ would be the mean time

Seminar of Andrei @ UTD: Third session

Meisam Hejazinia

04/04/2013

Only when the close solution is not available then you need to go over simulation, and markov chain.

How many times you want to draw is a questions to choose the sample size.

Significant test, 95% interval. 2.5 on each edge of the distribution.

Markov chains should onverge at some point.

Vectors of parameters should converge.

Full distribution will be separated to parts, and you want to make sure that all the conditional distributions should convrge.

If you can do it directly then the markove chain is not needed.

Split to chunks, and you separate and condition them. Split the distribution into conditional joint distributions.

We can simulate from full distribution directly. You draw from conditional, and marginals and then you would be able to simulate them.

We can spli $p(a, b, c|y) = p(a|y)p(b|a, y)p(c|a, b, y)$.

you can always create $p(a|b, c, y), p(b|a, c, y), p(c|a, b, c)$ you can draw from this. You need to be careful about where the point is.

States for above would be the value of (a, b, c) at each of the times. you condition on the previous times value of these states, and as you condition them you will get new values, and as each of the values as converge when we have on the value of states with regard to the previous time value then we say it is converged.

To calculate the likelihood we multiply everything.

$p(\mu, \sigma^2) = p(\mu).p(\sigma^2)$ then $p(\mu) \propto 1$, $p(\sigma^2) \propto \sigma^{-2}$. $p(ln(\sigma^2)) \propto 1 \Rightarrow p(\sigma^2)$, yout take the jacobian, and it is not about jeffery, since that would need information matrix. Any function could be proportioniate to 1, so this is not the problem.

To simulate first you draw from μ, and then take from σ, and then plug in from them. As a result first you take the marginals.

The main point that we try to calculate this is that we are not able to calculate the Gamma.

Any CDF is normal distribution between zero and one, so you use cdf^{-1} to use for distributing the random number with specific distribution.

As long as you know the distribution the method would be what we do here.

Simulate the data, compre distributions.

To check whether the model and simulation worked, you abstract model from the data, and you generate the outcome and compare with real data.

You a researcher must improve the model, and you should show that you captured the dynamic that previously has not been captured.

Truncated uniform would take considerable processing power.

Non parametric: no assumption of the distribution. Each data item will have its own condtribution to the frequency.

Rejection sampling. Algorithm: draw θ at random from probability density proportional $g(\theta)$, and accept θ as a draw from $p(\theta|y)$ with probability $p(\theta|y)/(Mg(\theta))$, otherwise reject and return to previous step. For truncated normal you remove things that are outside the value.

You just look at the part that has μ, you can use this for markov chain simulation, and for the inverse gamma distribution for calculating from joint distribution, you just compare the forms.
everything would be dependent upon the previous up to that point in markov chain.

chalesci decomposition of matrix

Homework: Gibbs sampler Gaussian linear regression with semi conjugate prior, derive gibbs sampler algorithm. Do it for yourself.

$y_t = x_t\beta + \epsilon_t$

$t = 1, , n$

$N(p, \sigma^2)$

$p(\sigma^2, \beta|y)$

$p(\sigma^2, \beta) = p(\sigma^2).p(\beta)$

Semi conjugate. The prior should look like. Prior is conjugate so.

$= (\sigma^2\beta)^T$

Derieve the following:

$p(\beta|\sigma^2, y)$

$p(\sigma^2|\beta, y)$

Try your best effor to do it, on your own, so that you can internalize it.

Both of them have different distributions, yet conjugate

Seminar of Andrei @ UTD: Forth session

Meisam Hejazinia

02/11/2013

In bayesian you are interested in the whole distribution and not just estimate. This is why we worked over both β, and σ^2, and we tried to estimate both at the same time.

You can normalize it later to make sure that the area is one. You just here try to find the proportion.

This is called Gibbs sampler, and you will find about it when we talked about metrapolist hasting algorithm.

You will look at the relation, and remove anything that is not canceled out, mean, is function of the main.

If priors not related through the sigma square it would be very difficult.

Semi conjugate since beta and sigma square are independent. There are situations that they are not independent, but unless you have reason that they are dependent do not relate to it.

You don't need to use this method mechanically.

The simplest way to simulate is to simulate it directly.

Things can go wrong anyway, and you just make sure you don't give up.

Do not put too much time in finding out about prior.

When you tighten the range of variance you may need to run many simulations to converge to your number.

Moments could be compared for checking whether the result is the same or not?

Multicollinearity should be handled by frequentist approach.

The first step is to take the parameters from data and develop model for random x, random y, and check whether the code works.

Second step is to take x from data and simulate y based on your code, and then check whether the real y is compared with the y simulated.

Comparing simulated y and real y could be done through histogram, or through checking the moments.

informal report.

HW Check this model over your own data.

what priors? how simulatation is sensitive to them?

Remove assumption and check how the simulation will look like.

Frequentist integrate beta and sigma, here is simulation based.

Based on the output that we took from the code

we take today.

Since data is i.i.d. once we get β, σ^2 then we can put $(\tilde{y}|\beta, \sigma^2, y)$, is the same as $(\tilde{y}|\beta, \sigma^2)$.

To simulate $p(\tilde{y}|y)$ then you just draw data from $(\tilde{y}|\beta, \sigma^2)$, and you have already did for $p(\beta, \sigma^2|y)$, and then integrate over them.

$p(\tilde{y}|y)$ is extrapulation, so you may be able to calculate it from frequentist approach as well.

This is as the process we said above, except we do it multiple times.

second homework

print out last page of the the book.
200 coin flips, 115 tails.

$H_0 : p = 0.5$

$H_1 : p <> 0.5$

Frequentist:

Multiply the probabilities and take log.

Emily estimate $p * 115/200$

You take ratio and say that it is not right.

Bays factor:

Not fair to compare model since one has free parameter, so it takes the integration.

Last slides contains two papers, you need to review.

Pay more attention to the first one (chip).

You can use bays factor to find determination.

This is one way to show that your model is better than others.

Seminar of Andrei @ UTD: Fifth session

Meisam Hejazinia

02/18/2013

18th of March would be the time for the midterm.

The OLS and the Bayesian should have the same estimate, so I need to check why mine is different.

Small number of parameters, it would be a problem.

We will start with the probit model and then we will deal with marginal likelihood.

Next time we will talk about meteropolis hasting, and markov chain will be discussed there.

When we don't observe full variable.

Utility some function of the price.

You need to ask everybody for utility. The person using the product does not know what utility is. The utility of buying something is greater than not buying.

If I decide to buy something means, my utility of buying it would be greater.

Two possibilities: buy or not.

Multinomial and multivariate will have the same kind of logic.

H is link function between the probability of something happening, could be anything. Usually it is used logit, and here it is probit.

Logit has fatter tail, and normal distribution of probit has thinner tail.

Multinomial, blue buss, and red bus. There is connection between them. Student t has fatter tail. The parameters will be distinguished. Theoretically they may be different due to the shape.

Probit has much more complicated. IIA in logit. 15 variate choice, and reasonabily large number, there is no computer can evaluate it.

Logit is simple to write the likelihood and estimate it.

Bayesian exact opposite is true. Logit does not have the family. Meteropolis hasting would be used for the logit. Student t would be different, yet logit would be different.

Multinomial problem you need to think differently. There are different ways to deal with mixed logit, but when you use probit there wouldn't be problem.

You try to get the best guess, to understand what underlying utility is.

Job candidate can get the utility based on analysis.

The main point is to recover complex structure.

Trick: connect data that you have Y, which could be 0 or 1.

When you take the difference it is like you get the

result of comparison.

We have 15 choices, and you can recover the number of choices. It is similar to what you have in the frequentist.

Introducing Z can help you to do the regression. Ultimately you don't know what Z is.

Homework code indirectly to find how it will work. We will simulate from normal distribution until we get draw from main.

It is markov chain, so it should always be conditional on all the current values.

Homework work over fixing the code.

For bayesian people usually use R programming.

In hypothesis testing. It is for model selection. What generated this data. You want to know whether this data came from this assumption or another.

Likelihood would be $\frac{115}{200}$

$exp(gammaln(201) - gammaln(116) - gammaln(86) + 115 * log(p) + 85 * log(1 - p))$ for the calculation of the value of likelihood difference.

If we set p to 0.5 and 115/200 then we will get two different value of 0.006 on one side and on the other hand 0.5750 which tells you that this could not be really could not be 0.5.

You can not compare, and you can not say it is the best. You can not adjust for it.

We forced it to be 0.5, since it pushes it to go to the different direction, since it is the combination of outcomes.

Bayesian tried to integrate them out.

We try to get non parametric model as we integrate them.

To calculate the bayesian you need to integrate the beta, and you will take the following.

it would be $\gamma(202)/(\gamma(116).\gamma(86))$ and you then you convert to factorial to find the number.

Since we know that integration of beta would be one, so we used beta.

This integration will be equal to the ratio that we calculated mean 0.006/0.5750 the first one does not have any parameter.

Confidence interval.

Decisive mean we can reject the null. Shows strong confidence that this model is better.

Homework: calculate the marginal likelihood for your regression model.

We need to be careful about how we select the priors, since i getting ln we would have problem since it will go to infinity for small avlues.

Homework: Linear regression, binary probit, and other model, try to run on your models.

Three people should present their findings.

Seminar of Andrei @ UTD: Sixth session

Meisam Hejazinia

02/25/2013

Everything conditional on the most recent value.

In the previous code we condition on $z0$, so you need to replace $z1$ with $z0$.

Burn in period is used sometimes.

You need to check whether all the parameters converged or not.

Burn in period is to make sure that it converges.

You can look at the graph instead.

Today we will talk about meteropolis hasting.

Marginal likelihood tells us how to select the model.

On bayesian you would not be able to through everything, and you just should think about whatever covarients make sense.

When you have many different variables, factor analysis could be proper method.

Every source is unique, and you can not use the same code for everything.

R− has building block built in and you can use to do the updating.

The week after the spring break you need to provide the powerpoint slides showing your data, what you are planning to do, and the model you will use.

Heterogeneity could be included in the model in the form of bimodal priors.

Poisson model for quantity, and the number of products to each category bought.

Multinomial choice.

Charactersitic of optimal pricing policy, when and how much to do the mark down.

Generate multinomial choice model, and aggregate them. You need to make sure that the result would be unique.

Characteristic of the product could give ou what would be up for most of the time.

The optimal mark down policy for specific group.

On march 25 you need to present your papers in the same way.

You should not have the prior so that it would be the only determining factor.

MCMC Simulation.

The paper of green and Chib. Read the paper.

Transition probability. Two state between which you can go.

The only thing that matters is transition matrix.

$P(x, A)$ in this paper means $P(A|x)$.

Borel set.

$p(x, R^d) = 1$.

You are searching for stead state distribution.

Metrapolis hasting it puts is in reverse. We know the distribution of final result. The question is where to go that finally distribution converges.

invariant distribution means converging.

the function should be reversible means $\pi(x)p(x, y) = \pi(y)p(x, y)$

It is like Gibbs sampler that you draw, but like accept rejection algorithm you remain on the current state. As a result it is something in between.

$\alpha(x, y)$ would be the probability of accepting or rejecting.

In the Gibbs sampling, the probability of result will have the same density as the probability of the previous tate, so it is the special case of $\alpha(.) = 1$.

Metrapolist hasting is more efficient than Gibbs sampling.

One way would be random variable, and each condtion on the previous one. The convergence for Gibbs was instataneous, but here for Metrapolist hasting could take a long time.

Binary Logit will be done today for coding.

We start with random walk, and we decide where to go conditional on starting value.

The density of logit, and proposal density multivariate normal.

The likelihood would be logit here.

$p(y_t|\beta) = (\frac{e^{x_t\beta}}{1+e^{x_t\beta}})^{y_t}(\frac{1}{1+e^{x_t\beta}})^{1-y_t}$

$p(y|\beta) = \pi_t p(y_t|\beta)$

$p(\beta) \propto 1$

$p(\beta|y) \propto p(y|\beta)$

$logp(\beta|\beta) = \Pi_{i=1}^{n}(e^{x_t\beta})^{y_t}/(1 + e^{x_t\beta}) = \sum_{t=1}^{N} y_t log(e^{x_t\beta}) - log(1 + e^{x+\beta})$

With probit you can do gibbs sampler, with logit you can not.

Read the paper of today, and new paper. Read them carefully, so that in case any question arises, we resolve it carefully.