EM114026大数据计量经济分析2025年10月14日第7周星期二第7-8节-pptVideo\_原文

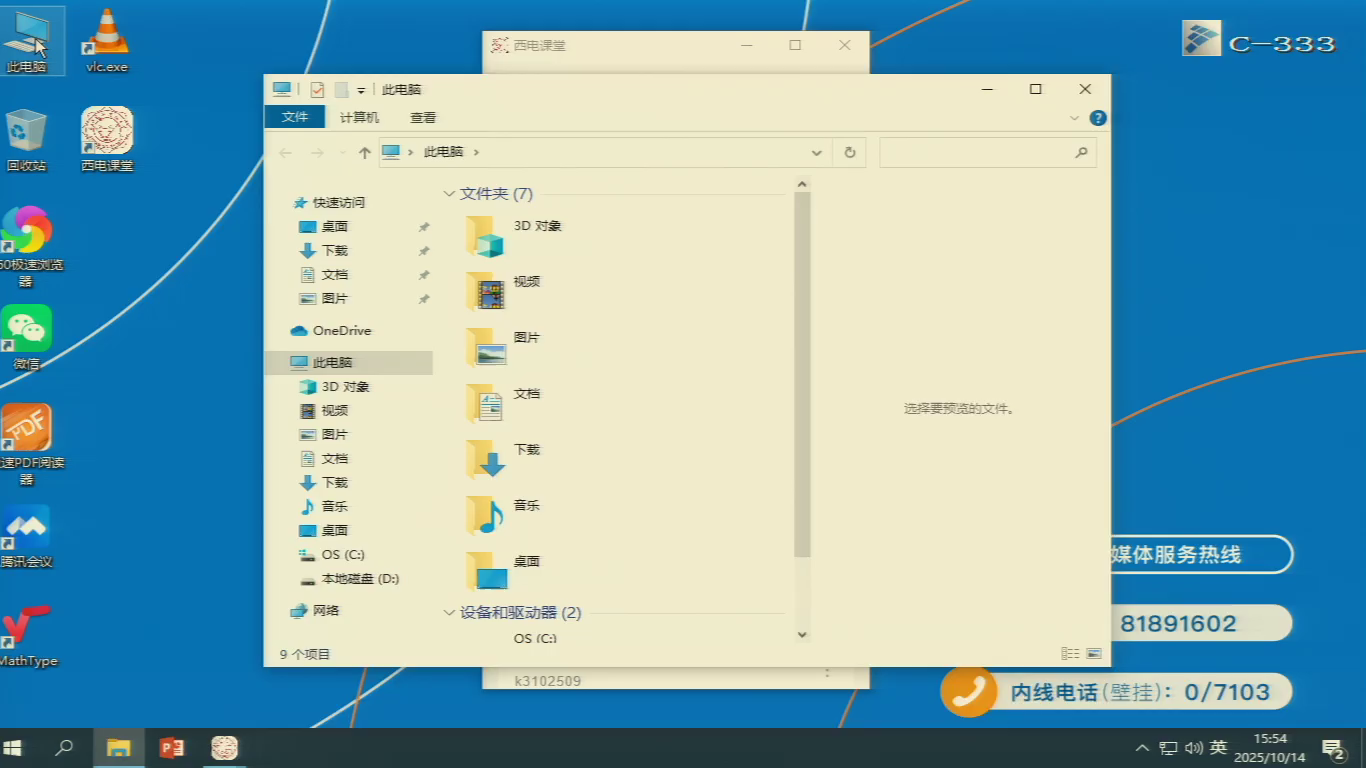
2026年01月04日 13:05

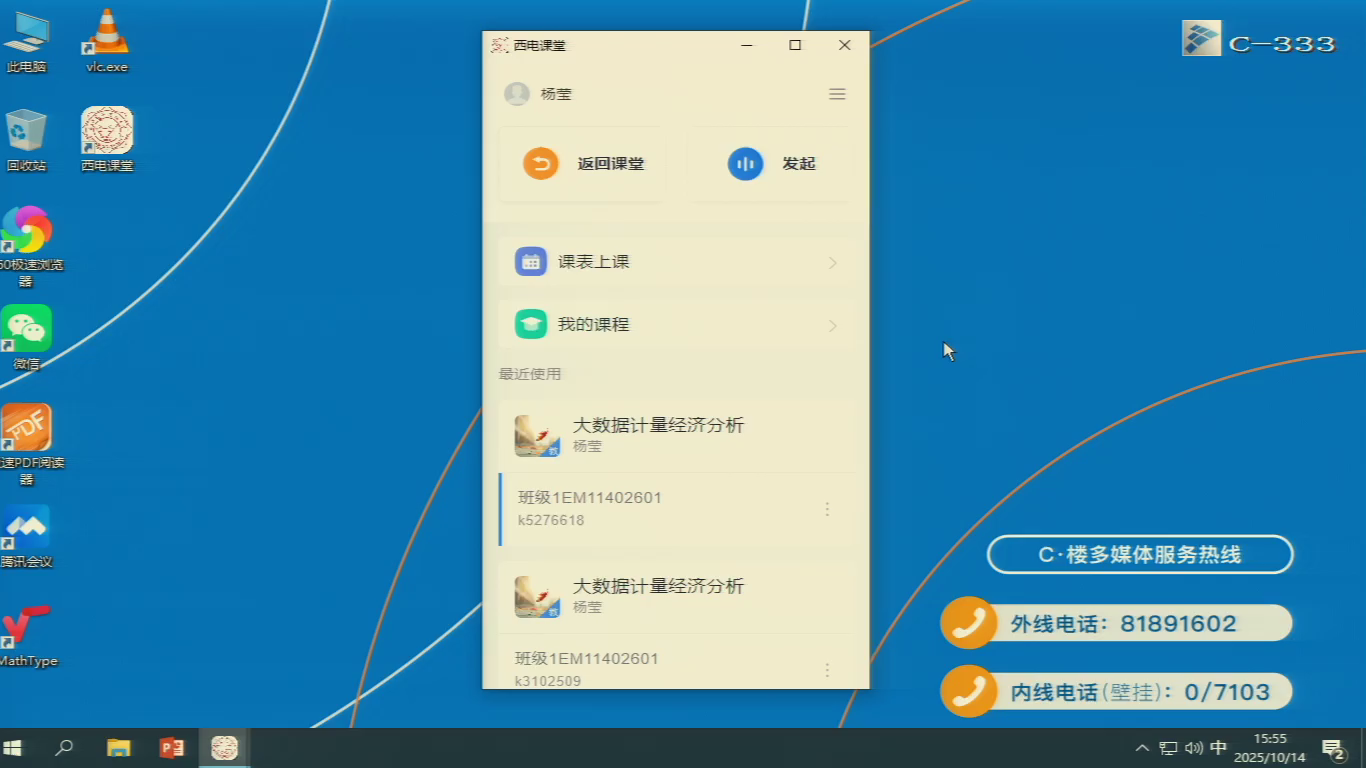












发言人 05:52

好的，我们这节课继续我们这个大数据计量经济分析的这样一门课程的学习。在上一节课我们给大家讲述了前面这个什么计量经济的发展起源和计量经济分析的步骤。我们在开始这门课这一节课学习之前，我们一起来回忆一下我们上节课的内容。我们找个同学来看一看，我们这个计量经济分析它有哪几个步骤。

发言人 06:36

杨家豪。

发言人 06:44

能回忆的起来吗？你可能不一定说的准确，那你大概记得多少收多少也可以。我们要做这样一个量化分析，首先要看。应该是要那同学下的对，明确你要到底要干嘛，你的这个问题是什么？之后因为我们是一个什么计量经济，或者说计量的分析，它也是个量化的分析，所以我们这个问题的分析建立在什么东西的基础上？

发言人 07:30

数据和什么模型？就是我们有了数据，有了模型才可以开展我们的分析，那我把这个数据假设说什么Y等于A加BX1加BX2加B2X2加6，我这个模型确定好了，X1X2Y的数据我也有了，那我下一步要干嘛呢？已经开始我还不知道什么这个模型是因为我知道它是什么，Y等于A加B1X1加B2X2加6。这个时候我还有一个问题不知道是什么，虽然我有数据了，但是我这个模型里面有AB1B2我不知道对吧？那我就要把数据带到这个模型来干嘛？估计求解？求解这个A和B这样参数的值，这个时候我才知道这个什么模型的具体形式，然后才是什么检验吧？刚刚讲的，然后检验之后，假设说我们现在得到的这个码AB1B2这些参数都是可靠的，或者说通过检验了，然后再下一步是什么？

发言人 08:57

对，可以得出结论，我们可以用来什么评估政策，或者是拿来预测，对吧？好，请坐。这个就是我们整个计量分析这一门课程里面的一个重点。因为我们在之后给大家讲解也会，尤其最重要我们上节课也强调我们这门课程最重要的是什么？就是估计出我们这个参数出来。因为什么？你设定模型其实很简单，我可以随意的设定，但是你这个模型设定的对不对，也会影响到你估计出来的参数，你设定的这个模型是啊指数的、对数的、线性的，或者说其他更复杂的一些形式。

发言人 09:44

这个是基于什么？基于你的这个理论认识，还有什么你的这个数据，他们之间的这个散点图的关系。比如你画一个XX1和Y之间的关系，你发现它是一个指数类型的这种关系。那模型设定我们就设定成一个指数对吧？

发言人 10:05

好，这个是啊我们上一节课给大家讲的这样一个具体的这样一个分析过程。从这样一个图表就可以看到怎么样我们一步一步的去最终把我们这个什么量化分析的结果拿来进行一些相应的应用。在这一节课，包括我们还给大家讲了我们怎么样去获取数据，从哪些渠道可以获取出，或者数据分为有一些类型，我们也找一个同学来回忆一下。

发言人 10:45

王红祥。好的，好点了。那我们这个数据有哪样一些不同的形式呢？

发言人 11:08

对面板数据是不是一种形式？还有一种是什么？比如说我们按GDP从2000年到现在，每一年一个GDP对时间序列的数据，这个是什么？纵向数据？也就是它是随时间变化的那还有一种，我现在把我们班同学每一个人的收入情况都都收集一遍，统计一遍。那这个数据是什么数据呢？但是他也是纵向的数据吗？不是？它是一个什么横向，比如说横截面的数据？我在一个时间点收集很多对象的这个数据。

发言人 11:55

也有一种，我从2000年到现在，假设说我追踪调查了我们班每个同学每一年消费支出的情况。那这个就是一个什么？这是一个混合的面板的数据对吧？就是有每一个人每年你就可以想象到，它是一个什么面板数据。好，请坐。这几种形式就是我们主要在建模构建模型的时候会遇到的那在这一节课我们继续给大家讲述，在大数据时代背景下，我们怎么样开展这样一个计量分析，或者说会存在哪样一些不同。

发言人 12:37

大数据它有一些特征？它的有很多人对这个大数据给予相应的定义。有的人认为这个大数据意味着维度度，或者说它的什么。也有的人认为它是规模大，就是它数据量比较大。

发言人 12:59

在这样一个大数据时代，它主要是什么？我们这个互联网迅猛发展导致的那我们国家的网民，他这个数课本上的数据有点滞后了。那现在？已经超过十亿多。那普及率也是互联网普及率也是非常高。尤其是现在手机联网对吧？手机用户基本上都达到90%。这样大规模的数据我们他还有一些什么来源广泛、实时性、价值密度低这样一些特点。

发言人 13:41

与大数据对应的，也存在一个小数据的这样一个概念。小数据它是指什么？一般来讲它就是什么？体量比较小，实时性比较低，通常都是无线采集的数据。比如说我们就统计或者抽样一下西电学生的这个消费的情况。那这样是不是就可能数据量比较小，对吧？我一个学年抽取60个学生，这个就是一个小数据的这样一个特征。

发言人 14:15

大多它来源于什么？调查问卷的数据以及企业运营的数据，或者说小数据，它基本来源单一的数据。比如说我们国家统计局公布的这样一些经济指标的数据，或者说是企业财务系统里面的数。那大数据它倾向于采集多种的，然后大多是来自于网络。

发言人 14:41

另一方面，小数据它基本以什么结构化的数据为。比如说我们只有什么一些数值型的数据，那大数据它就涵盖了非常多这样一些非结构化的数据。比如说图片的数据，客服语音留言、网站日志。比如说现在很多人做一些什么直播这样一些的分析，那他很多时候会买什么？这个主播他在整个直播过程中，他的语音，包括什么时候会把这个视频呈现的人员数量或者他们情节的互动这样一些这样一些视频的数据，图片的数据以及语音的数据都把它摘出来，作为非结构化的数据展开分析。那大数据是不是都是好的，或者是小数据是不是都是坏的呢？其实不一定，他还是根据什么，根据你的研究问题和你的问题导向来说，一般来说我们倾向这个数据量大一些，丰富一些，这样你去捕获的信息就会比较多。

发言人 16:00

刚刚我们讲到大数据，它有一个特征，它的价值密度比较低。就是说大数据虽然很多时候还是一些实时的数据，规模也很庞大，但是它也会蕴含什么很多的冗余的信息。就有的时候你要分析这个大数据的时候，或者这个大数据可能有的并不一定比这个小数据要好。一般来讲，小数据它都比较准确，然后信息含金量比较高，分析它的成本比较低。那大数据它的分析成本就比较高，很多时候你要在很大规模的数据里面去找到你想要的信息。此外大数据它涉及的维度也比较多，很多时候我们相解决相关性的问题。此外什么大跟小很多时候是相对的。小数据也有可能会有一些什么比较大的这样一个价值。

发言人 17:07

那大数据它对这个计量建模也会带来一些什么挑战？首先一个就是大数据它因为形式的这样一个多元化，因为它比如说包含什么音频的、视频的这样一些数据的时候它会有一些比较复杂的这样一个影响在里面。另外传统的这样一个结构化的数据它从简单的这样一个点数据扩展到区间数据或者说符号数据、函数型数据，它也会对计量带来一些影响。那怎么样对这些信息含量丰富，数据形式多层次的这样一些区间数据，符号数据和函数型的数据进行计量建模？就是现在具有计量方面具有一些比较挑战性的这样一些的研究工作。

发言人 18:07

此外数据变量它的维度比较高，也会给西藏建模带来一些。比如说在传统的这样实证研究当中，我们这个建模都是线性的形式变量，我们就根据这个理论或者经验来进行选择。但是大数据大家可以想象到，我们拥有海量的信息的同时在我们给定的样本中，可以进行分析的变量维度就会更高，甚至会出现远高于样本量的情况。这样的数据我们把它称为什么胖大海。在高维数据当中，我们怎么样或者说如何有效去筛选信息，就变成了大数据既然面临的一个挑战。这是什么意思呢？

发言人 18:57

就从前比如说我们要调查我们班同学这个卡卡消费支出情况的话，我直接就通过问卷调查去去收集大家的这样一个相应的指标信息，对吧？比如说我收集的就一个变量，你请问你每每个月的消费支出是多少元？这个数值是不是就一个变量刻画出来了。但是比如说现在学校或者说我们可以通过跟银行合作，能调取到大家实时的这样一个消费信息的时候，大家可以想象到你的这个消费行为，我是不是就有很多的变量可以去刻画。

发言人 19:42

除了时间维度上的，我可以把你的什么你的消费区间刻画出来。你每天的消费，每周的消费，每个月的消费，还有什么你的消费变化，还有什么你的这个相对的消费量，或者说你有一个具体项目上的这样一个消费。比如说你的食品消费，你的什么衣服消费，你的剪头发消费。就是你的很多细节性的信息，甚至它实时的信息都给你捕捉下来的时候，你怎么样在这样一个问题里去筛选，或者说是说选取的这个变量，去刻画你想研究的这样一个关键的问题。此外这个变量的关系也更加复杂，会给什么计量建模带来一些相应的挑战。

发言人 20:38

此外还有一些什么网络数据的分析。比如说我们在做什么消费行为这样一个分析里面，他也有可能不同同学之间的什么消费行为可能也会相互影响。这个时候我们可以从银行的数据里面，比如说大家成为了朋友，你可能什么去聚会，这个密度就变高了。那你AA的数据，AA消费的那个数据它也能捕捉到的时候，这个时候我们就可以把把消费行为刻画成一个什么网络性的消费行为。这样的一些数据成为网络数据的分析。下面我们也一起来看这个机器学习、统计学和计量经济学的这样一个关系。在之前的课程当中，我们是不是给大家讲过我们这门计量经济学和哪两门课程，哪几门哪三门这个课程的关系是最大的。

发言人 21:55

我们也找个人一起来看秦雪婷。

发言人 22:05

我们这门课程是不是之前给大家画了一个图？三个圆圈交集出来是我们这门计量经济学的课程。在之前课程里我们给大家讲过它是哪三门课程的什么结合体或者说交集呢？

发言人 22:31

他自己给自己教对经济学。还有。统计学还有什么？

发言人 22:45

还有数学对吧？对，好，请坐。之前我没有给大家讲到什么计算机算的课程。随着近来的什么计算机这样一门学科的发展，或者说人工智能这样一门课程这样一个发展，也就是所谓的机器学习，它的什么应用越来越广泛。

发言人 23:13

基因学习，我们可以想象到它跟统计学的关系是很紧密的对吧？比如说大家学学到的，不知道大家学过没有？比如说我们这个神经网络，或者说像很流行的什么大语言模型，它是由什么transform这样一个结构。它底层来讲它都是什么统计学的这样一些内容，对吧？

发言人 23:36

那他跟计量经济学之间是什么样的关系呢？大家可以看到经济学习它实际上也是用数据去喂出来的，或者用数据训练出来的对吧？它通过各种算法，比如说他直接聚类分类，我直接我不会去考量什么，你不同变量之间的关系，我只是什么，我把所有的信息或所有相关的这样一些数据喂进去。最后通过一些算法，比如说你要什么也就是最小，或者说你采取一些相应的设定形式去生成这个什么神经网络的节点，然后去确定出来怎么样能得到一个最优的结果。

发言人 24:27

这个时候机器学习很多时候他它就相当于是什么从上或者从数据中去学习怎么样去完成任务。他跟我们什么计量经济学最大的区别是什么？我计量经济学这门课程会重视什么因果推断，或者说重视变量之间的互相的影响关系。就是我们需要得到X怎么样影响到Y的那机器学习它无所谓。你可能比如说有一些Y变量的分支，你也可以把它放到这个X左端就是你导出或者预测出最优Y的时候，他并不重视什么他们之间的因果关系，他只重视你只要能预测出下一期，或者说预测出有良好表现的这样一个Y那我就认为你这个继续学习这个方法是可那你说他会用在什么最优化和算法这样一些形式里面的统计学，它作为底层的这样一个数学的比较偏数学的这样一个课程，还是比较偏理论性的。因为他主要的方法，比如说是鉴定理论。他更多的是在可解释性方面有很强的要求。同时他也要用在把预测和因果推断。

发言人 26:08

计量经济学相较于其他两门学科，它是什么？大家可以看到我们经济学最主要的是什么？我们要去解释相应的经济现象。所以去是啊最终是什么变量之间的这样一个因果关系，或者说因果推断的。在方法论上的机器学习，它主要用把最优化的方法表现为最小化目标，获得损失函数。他们之间的关系，大家通过这个表就可以很明显的去看出来。

发言人 26:50

最后在这些导论课当中，也会给大家介绍什么，是我们现在这个计量定义分析里一些主流的软件。在开始给大家稍微介绍一下这个reviews。Reviews通常来讲就是如果大家学习统初级的统计学或者初级的计量经济学，一般都是用这个muse软件。但是现在因为可能很多本科生他也会有什么做科研，写文章这些需求。那refuse，现在很多时候它不适用。所以在我们这节课程当中，我们给大家用的软件是西塔这样一个软件。

发言人 27:34

那同样的我们也需要去了解一些，还有一些什么其他的软件，我们可以选择的来看一下这个US它是什么？它是一个比较简单的，可以通过一些什么可视化的。也就说通过这个界面点击的方式，你就可以来进行相应的一些分析。

发言人 27:57

Refuse它其实也不是说都不可取，它比较擅长的方面是什么？它比较擅长在这个时间序列的回归分析单位跟检验及格兰杰因果检验协调检查ramm都二维码模型就是最经典的或者说最好用的这样一个什么预测的模型。比如说大家想要预测什么金价的走势，或者说要做一些时间岁月的分析的时候，用这个reviews的软件。可能比如说大家看个教程，半个小时1个小时就学会了。因为它是一些比较简单的点击式的操作。你把数据导入到这个软件里面，只要点击这个相应的回归分析，它就会出来结果。但是你对于数据怎么样分析，怎么样建模，这些底层的知识也是需要具备的。所以这是我们这节课通过这门课程给大家开展。

发言人 29:02

那除了in use，还有这个软件叫什么SAS这个软件它也是一个什么比较专业的这样一个统计分析和数据分析的软件。他在这个什么大规模的数据分析里面，也是非常好用的这样一个它功能包括什还可以很好的去对接。数据库进行数据访问、数据存储、应用开发、运筹外包编制等等这样一些数据分析的方法。

发言人 29:44

SAS它是一个什么商业的软件？一般它的什么安全性性能比较高。也就是说这个软件通常在什么公司里面应用比较多。因为它是一个比较还是一个比较昂贵的这样一个商业性的软件。就是它的功能性来讲是非常丰富的那同时，它也是一个收费的这样一个软件，他在银行或者说其他一些咨询公司里面，用这样一个软件会比较多。那西塔，它也是一个收费的这样的软件，但是它的这个码价格会低一点。

发言人 30:28

用SAS这样一个商业软件，通常是基于安全性的考虑。比如说你银行的数据，你要去调取什么国家网网信办、经济办这样一个数据接口的时候，他很多时候是考虑什么安全性的角度出发，而不是功能性。那data软件它是一个什么比较功能强大，简单也是比较简单易操作。它可以有不同的形式，就是在一些简单操作的时候，你可以去点击式的操作它的界面。除了点击式的操作界面，如果说是有一些复杂性的操作和数据分析的需求的时候，你也可以通过一些代码的形式去保存在读文件里面，就可以一次性的对数据进行相应的计算运算、数据分析、数据检验。

发言人 31:27

那很多的这样一个数据库，它公开的数据，也是date格式的这样一个数据。比如说北大他们有这样一个今天给大家讲过，他有一些家庭的追踪调查的数据，他就是用data的格式的那这个数据，这个软件，在什么我们现在用这个什么来进行科学研究，或者说是计量分析是最主要。就大家可以看到一些主流期刊上发表的文章。如果他说他构建的计量模型来进行分析的话，大部分都是用这个data软件去去进行实现的那同样私下它也是一个收费的软件，但是它价格的相相较于SAI是要便宜很多。可能我们从前在交大的时候买过，可能就比如说一年2万块钱，比如说你这个个人账户他就可以购买的那相较于前面这样一个。

发言人 32:38

这样一个软件R它就是基于什么开源的。就现在非常流行的开源的、开放的，大家都可以共建的这样一个完全free的这样一个软件。它的功能也是比较强大的那之前我给苏志生讲过这样一门儿语言的课程。我们从前也是会用这个R开展一些什么相应的数据分析或者计量建模的这样一些内容。而相较于其他的软件，它有一个很大的优势，它的这个实时性会非常强。因为它方法的更新，会马上大家开源到这个R软件里面，然后实实现这个这样实现R包的调用，你就可以用最新的方法来对数据进行分析。

发言人 33:39

这个是啊作为一个开源软件两大好处之一。一是它这个实时性，你的这个可获取的资源非常丰富。我觉得很多人会在这个R软件上面去进行贡献。就是你写好的这个R包可以共享到他这个平台上面，然后别人可以用你优化过的这个方法，或者说优化过的那个代码去实现一些数据分析的功能。

发言人 34:09

第二大好处，可显而易见，它是一个免费的软件，对吧？就是对于科研学者来讲，它就是一个最好的软件了。这个软件它的起源也是什么生物学家他们在做一些大数据分析的时候，他就去也是比较窘迫。所以搞科研都比较穷。他就没钱去购买这些商业的软件去进行数据分析。所以他索性就自己开发了一个，然后也作为一个免费的面向大众，所以慢慢的这个软件就也翻的那这个是我们这节课程会用到的教材和参考书，上节课也让大家去购买。一个是我们这个教材，就是大数据计量经济分析勤勤满月的这个，那同样我们也有两本参考书目，李子奈清华大学李子奈的这个计量经济学的内容。还有就是跟西塔相关的。

发言人 35:18

这个就是很多大家到了这什么研究生阶段，就相当于是一本代购的工具书了，就是计算机解析及data的应用，它会在这本书目里面去给你详细的阐释每一种模型怎么样来进行C的操作，然后数据相应的一些设定。好，我们这一节或者说这一本课程当中，导论课的学习我们就到此结束。主要的学习要点我们一起来说一下。首先我给大家讲一下什么计量经济学相应的性质，计量经济学其他学科的这样一个联系和区别，你学习计量经济学一些现实的必要性。其次我们给大家讲的就把降低学基本的思路步骤，以及我们这个什么模型设定、参数估计、模型检验的一些要求，以及还有什么模型中的变量以及相应的类型，让我们一起来回忆一下模型当中它有哪样一些变量和类型。

发言人 36:49

王梦瑶。我们这个计量分析的模型里面有哪些变量呢？或者说有哪些类型的变量？

发言人 37:12

这是我的模型，真的吧？

发言人 37:21

这是我们计量经济的逻辑最简单的代表。那变量有哪些？这个Y是变量的。

发言人 37:37

Y是什么因变量？那X是什么呢？自变量？所以他有自变量也有因变量？那这里面我们的BU是什么呢？误差项对吧？好，请坐。这个就是我们模型当中会涉及到电量以及类型。数据的类型我们也相应回顾过了。

发言人 38:06

参数估计的方法它也存在一些不同的类型，对吧？比如说什么最小二乘法OLS。叫什么简单最小二乘法，然后还有什么据估计广义最小二乘法，三阶两阶段最小二成长，三阶段最小二成长。这些都是根据模型不同的形态，我们要用相应的估计方法对这个模型进行估计，以及我们就是把建立计量模型的依据，我们计量建立计量模型的依据是什么呢？

发言人 38:47

黄诗怡。计算机模型。比如说我为什么能建立一个A加BX而不是A加BBO加B。就我怎么样来构建这样一个计量经济模型，依据有什么？营养关系还有那为什么这里不是一个A减BX你可以减B老板，是的。

发言人 39:35

就是X和Y的关系有没有可以由什么来界定？那么由经济学的理论来界定？就是假使我们这个收入消费的问题，现在真的得到了一个互相的关系。但是我们构建理论模型的时候，我们要依据什么这样一个经济学原理？除了经济学原理还有什么我们的实践经验？就有可能还能构建成什么理论，你能明显的知道什么这个收入分成，他的消费行为也不一样，所以他很有可能买A1加B1X加谬。然后当什么X假设X小于什么2000的时候，它是一种消费行为习惯，或者说一种消费模式。我们可以用这个模型A1加B一来刻画。

发言人 40:42

当这个什么收入。消费比较高的时候，它有可能？就是另外一种消费行为习惯了？

发言人 40:57

就是有可能分成的模型对分阶段的逻辑，有可能之间的现金的关系，也有可能是买指数的关系。这个要看我们经济学的理论还有什么实践的经验，对吧？就有可能你这个什么线性的关系是理论告诉你的那你在实践当中，你天天在这个综合楼转悠，或者你天天在这个什么学校转悠，天天看你看出来好像是不是你实践的经验告诉你，可能不同的这个消费型收入它有可能有不同的表现，你也可以把它设定成这样一个分阶段的东西，对吧？好，请坐。那也是说我们建立计量模型的依据来自什么理论，再加上什么实践的这样一个经验，这个是我们第一章的把这个学习内容。

发言人 42:05

第二张，我们给大家讲的这个尺码。怎么样来对大数据进行预处理？这个是我们这门特定的大数据计量的分析增加的内容。就是他跟李子奈课程或者教材里面不一样的地方。就传统或者说一些经典的计算机课程当中，它不会有这什么相应的数据的介绍。在我们这个第二章内容学习里面有大数据据处理，主要给大家会讲以下的内容。首先是把数据的采集，其实数据的质量，数据的预处理也会给大家讲一什么文本挖掘的这样一个实际案例。

发言人 42:54

首先我们一起来看一下数据采集的内容。数据采集它也分为不同的这样一个数据来，主要会存在这样一些类型。首先是什么业务的数据，比如说消费者的数据，客户关系的数据库存数据、账目数据等等。其次他也有一些行业的数据，比如说车流量的数据，PM2.5的数据。还有一些什么内容的数据，比如说应用的日志、电子文档、机器的数据、语音的数据、社交媒体的数据。这个内容数据现在应用分析的比较多的那第四还有是什么线上行为的数据。

发言人 43:52

比如说页面的数据、交互的数据、表单的数据、绘画的数据、反馈的数据等等。比如说你在这个购物网页上的浏览，比如说你打开一个浏览的数据，在哪个部分你鼠标停留的时间最长。他这些页面的数据或者说你这个人机交互的数据，它都可以捕捉下来。这样这也是一种线上行为的数据，还有线下的行为数据轨迹的数据，用户位置的这样一个数据等等。

发言人 44:35

那从我们这个什么数据获取的方式，主要来源于以下不同的这样一些系统。一个肯定就是企业的系统对吧？你很多时候你分析的数据都来源于这个码企业。第二是机器系统，比如说从设备里面对吧？比如说你的这个智能设备，你的手机里面提供掌握的一些信息的数据的。比如说像每个人有的，现在手机都可以记录大家在每一个软件上的停留使用的时间，对吧？

发言人 45:14

那第三是互联网系统，比如说包括电商的系统。你在抖音上面想要分析一些直播的数据，抖音的数据的话，它就有一些什么。比如说像陈妈妈，他有一些什么提供数据的平台，就是他把抖音的数据全部爬取下来，你就可以在妈妈这样一个平台上面，去获取到抖音用户的这样一些我的电商的这样一些行为数据。比如说哪个用大V账号或者按普通账号来带货，还有一个月销量是他带的产品是什么这样一个数据。同样还有一些什么服务行业，政府监管系统。

发言人 45:59

这样除此之外，还有什么社交系统的数据。比如说微信、QQ、微博、博客、新闻网站、朋友圈的。就像很多人会做一些什么舆情的分析，或者说是什么传播学的这样一些分析。比如说那个谣言它是怎么产生的，很多人就用什么微信的，有微博的这样一个数据来分析。也有人拿着什么微信的数据来。

发言人 46:28

比如说有一些什么内容看起来很荒唐，但是他却广为流传？比如说有一些自媒体标题大家写的不准，不是中国人，对吧？就有一些学者他们就拿这个作为研究课题，最后发现什么样的情况下什么一些荒诞的言论。所以这样一些什么不转不是中国人这样一些营销账号过去，然后最后实际上非常大规模的这样一个传播效应。

发言人 47:06

首先我们来看一下这个日志数据和网页数据的采集。首先是这个什么叫做日志的数据？它实际上是也是起源于互联网。很多数据比如公司业务平台，它会产生大量的日志数据。这些日志信息，它就可以看出很多有这样子的数据。尤其是对于外部的应用来讲。它一般来讲这个日志数据包括用户的访问地址，用户的购买数据或者用户的点击的地址。常用的这样一些日志数据采集，主要有浏览器日志采集和客户浏览数据。

发言人 47:51

老师的一个老师，他当时就是拿到了一笔这样一个日志的数据。这个可能大家都可以想象得到，这个日志数据是非常简单的一个数据。比如说他拿到的资产类似于这什么百度百就他拿到什么维基百科这样一个日志数据。他日志数据相当于把某个用户在某个时间节点，在什么维基百科上进行了几个动作。一个是什么增加内容，修改内容或者删除内容。就是这样一个非常简单的包括时间行为，就只有增加删除和修改。

发言人 48:34

拿着这样一笔数据之后，把这个什么日志的数据进行了一个挖掘。他最后发现他他从这样一个维基百科日志数据用户们来贡献的这样一个角度，把这笔数据拿来分析。最后发现的一款他从一个角度出发是他把这个维基百科当过程一个公共品，就把经济学的这个理论给加上了。最后这篇研究成果它在我们这个经济学最顶层的AER上面，除了它在AER上面也画在什么MS这样一个非常顶级的棋盘上面。虽然有的时候大家可以看到这个日志数据还是一个非常简单的。就是当时那个老师给我们讲，他那天研究的时候，他就说他拿到那边数据也是很头疼，可能每天打开看，每周打开看，他可能就这样看了三年，最后想到怎么样把这个数据的价值给它挖取出来。

发言人 49:43

其次是什么网页数据的采集。网页数据的采集大家可能都有所耳闻，或者说可能有的同学都能直接上手了。一般我们采集这样网页数据的时候，就看你这个什么技术的好坏，对吧？很多时候都是只要可见即可得。他可能不管他反爬虫多厉害，他只是说成本的问题，或者说是什么你的这个技术的问题。他一般是通过相应的这样一些数据采集器，来将什么网页展示的内容，或者一些没有展示的内容，它隐藏在这什么网页数据里面，都可以把它采集下来。那常用的一些或者说简单可以上手的这样一些网页采集的这样一些软件。你说有什么八爪鱼这样的产品形成什么，也是简单的点击式的操作，你就可以去爬取相应的网页信息了。

发言人 50:56

它一般都是什么？你输入相应的网页页面，然后它就会出现那个网页页面。你选取这个网页上你需要相应查取的信息，把它框起来。就是这个页面相应的内容，需要相应的指标，你需要你把它点击选取全取下来，它就会形成什么格式化的，或者说是结构化的这样一些数据形式。它有可能是文本，有可能是数值的。你再规定他的这个什么翻页的形式，就是你要采集的肯定不只是单个网页的内容，你要采取采采取的是，比如说你搜索某一个关键词，这个关键词涉及的所有的相应的信息点，设定好它的点击动作。那这样他就可以把这个相应的这个网页的数据给爬下来。有一些比较复杂的，它就需要通过什么python这样的软件去进行相应的数据爬取。

发言人 52:12

除了数据采集的内容，我们也要看一下数据质量怎么样来评判。数据质量最重要的就是数据的完整性。这个数据的完整性它是指什么？主要的信息不应该缺失，这个时候才能成为一个什么比较完整的这样一个相应的信息。一般来讲你可以通过什么你收集到的相应信息，看有没有一些什么缺失值。对这个数据的完整性进行评估。

发言人 52:55

很多时候你说大数据这样一个内容查取了什么几十万条的什么假设说是什么小红书上大家发帖的这样一些信息的话，那很多时候因为啥？比如说某个网页你爬的时候，他那个网卡了一下，它就会有一些缺失值出现。那这个时候，假设几十万条信息里面你是缺失了几十条。那这个时候这个数据相对来讲它还是一个什么比较高质量的数据。也就是说你要去评估这个数据完整性的时候，然后去兼顾它的数据规模，比如说你只盘下来100条，但是缺失了几十条信息，那这个时候这个数据就是不完整的对吧？

发言人 53:44

那有一些记录他在输入的时候，它也可能会存在一些你认为不重要，或者说理解错误的疏漏，或者说机器的故障存在不完整的数据。比如说什么传统的一手数据去收集的时候，你要把这什么调查问卷转化成什么数值型的对吧？你要看每每一个问题，这个人是打了五分，或者多少人给这个问题打了五分。那这个时候，你把把一份问卷转移到什么计算机上，它也有可能会什么输入错误，对吧？造成一些什么你说漏报的这样一个问题，这个是数据完整性方面的考虑。

发言人 54:33

除此之外还有正常数据一致性的考虑。也就是说我们收集到的数据，你需要去判断它的逻辑关系是否是正确和完善的那规范性它就比如说如果大家收集到大家什么身份证号码，它是一个规定的对吧？比如说它都是18位的，它都是什么七位数字，这个什么手机号码，还是13位。那这些规范性的数据，你就可以去检查它是不是存在一次性。

发言人 55:10

那逻辑它是指什么多项数据间存在固定的逻辑关系。最简单的比方来讲，假如说我们拿到的是哪一份销售的数据，你可以想象到什么单价乘以数量就是它的什么销售，对吧？那很多时候如果说你检查的数据一致性的时候可以低调。有的时候比如说他的最起码单价乘以数量，最后跟最起码你的记录的这个销售额不匹配，或者说再减去折扣的能量加起来它并不匹配。那这个时候它就会存在什么数据一致性的问题了，对吧？

发言人 55:53

那数据准确性还是指这个数据记录当中，他信息或者数据是否准确，是否存在异常。与他跟一次性其实还是比较接近的对吧？但是他也不一样的是什么准确性的问题。所以它不只是规则上的不一致，他这个什么也有可能什么人为输入的错误，对吧？有可能还会什么你在收集这个准确性的数据收集的比如说你在营业额上面可能小数点标错了，然后每一个营业额都加了一个，对吧？那这个时候它就是这个了数据准确性的这样的问题。

发言人 56:35

还有什么数据时效性的问题。比如说你在探究你的数据是否好是它的好坏的时候，你也要去考虑到什么它的时效性。比如说我们要探究欺骗学生收入消费的情况。我可能比如说拿什么20年前这个学校记录的学生的数据来分析，那可能就对我这个分析没有太大的用途，对吧？因为我去分析西电学生证收入消费的情况，我是想什么？比如说我作为这个学校的管理者，或者说我作为什么政府官员，我想要针对大学生出台一些刺激消费的政策，对吧？

发言人 57:21

那这个时候，我拿20年前的什么消费行为习惯数据来看。那这个时候这个数据可以说是好还是不好呢？不太好？虽然他没有缺失值，没有错误，然后数据也都是一致的。但是因为什么他没有时效性，对吧？他针对你当前的这个研究问题，它就不是一个好的数据。

发言人 57:49

同样的比如说这个什么？现在一些智能手机还有什么定位的功能，而且大部分开放什么实时的位置信息。这个时候就是什么你在进行一些营销活动的时候，你想要拿到这个码用户的定位信息。这个时候它这个数据就具有很强的什么时效性，对吧？

发言人 58:17

那现在有一些营销学研究的内容就是什么，他有相应的数据能捕获到。比如说你经营的商场之后，你在什么，比如说阿迪达斯和耐克这个店，假设说你进了，它能捕捉到你进了阿迪达斯的店之后，再到耐克的店，他可能马上能实时的给你的手机推送一张优惠券，对吧？那这个时候就会把大大刺激这个消费。那它有待于就有待于实时的把你的这个轨迹的数据，或者说你的这个数据的时效性。

发言人 58:59

还有在我们评估这个数据质量的时候，也需要去考虑到什么数据的这样一个相关性吧？那比如说我们在预测什么交通事故发生率的这样一个模型里面，这个时候大家可以想象到在这样一个交通事故发生，还有一些关键的数据信息。比如说什么驾驶员的年龄和性别对吧？那比如说这些信息可以获得的情况下，那这个数据可能就是比较这个模型就比较好，否则这个模型的精度就会受到限制，对吧？那也就说什么，我们这个数据质量还是要根据什么相应的问题去考量。

发言人 59:54

在特定的问题情况下，有可能比如说有一些必不可少的或者说关键的这样一些信息。比如说你在探究消费的这样一个问题的情况下，假设说你都没有大家的这个收入信息。那这个时候你就会想到这个数据质量它是一个什么？非常糟糕的一个质量，对吧？因为它是在这个消费问题里面，或者说在什么交通事故预测的这样一个问题，它是必不可缺必不可少的这样一些。这些变量关键变量的缺失，也会极大的去影响你这个数据质量。

发言人 01:00:35

好，假设说我们现在已经获取到了准确的、一致的、没有缺失的，然后关键的变量也都存在，而且是时效性很高的这样的数据的时候。我们怎么样来对或者说当某些情况下不满足之前的条件，我们来对什么数据怎么样来进行预处理，让它能达到或者说实现能用来我们这个计量分析建模的这样一个数合规的或者说是合格的这样一个数据。有一些数据，他的什么可能你的什么来源比较好，你的这个格式数据完整性、准确性上面也都比较好的，这个是非常理想的情况。那大家在现实的生活当中，或者是实践的应用当中，大多数情况下你拿到的数据都是啊不一致的，或者说有缺失的，或者说是不准确的。有时候会存在一些异常值、缺失值这样一些情况。那这个时候你就需要对数据进行一个预处理的这样一个工作。好，我们先休息一下。

发言人 01:19:58

好的，那我们继续我们这个计量机分析的课程。接着我们上节课讲这个实践当中，或者大家拿到手的数据，大多都是比较脏的。它有可能会存在什么缺失的数据，或者说是异常值等等一系列不一致，不准确等等一系列的问题。我们在日常把这个数据用到我们这个计量分析模型里面，我们首先要保证它不是一个咖啡机，对吧？就很多时候它是一个什么咖啡机，用咖啡机out这样一个问题，不管你这个模型设定的多么优美，多么准确，你的这个什么估计方法多么的科学。有的时候你的那个什么源头，你拿来分析这个数据本身就有问题。那这个时候你后面再再多努力也是功亏一篑。所以我们会拿一章节的内容来强调数据预处理或者说数据分析之前的这样一个步骤的重要性。

发言人 01:21:15

可能很多同学如果没学这个课程，可能也就以为他是一个什么样非常小的一这样一个部分的内容。实际上，它虽然可以讲的内容不多，但是他在这个什么计量分析，或者说大量的什么数据分析的过程当中，它其实是最重要的。有的时候他可能占据能占据你80%、90%的这样一个经历，在这样一个部分，虽然它只有一张纸，但是你实践去处理的时候，这个部分就是这么重要。

发言人 01:21:50

有的时候是什么？你在数据分析的时候，你假设说能发现或者说是能由现有的数据当中组成一个比较好的这样一个变量值数据形式的话。这个时候很多时候就决定了你的结果还是一个什么比较有价值的这样一个工作。

发言人 01:22:15

首先我们来看数据集合并的这样一个内容。数据集合并就是我们在大数据分析的时候，很多时候我们会拿一个什么多元的数据。有的时候你就要去把这个什么数据以及去进行相应的这样一个合并的处理。再去合并处理之前，我们要先分析或者说先观察数据变量，去仔细分析了每一个数据集变量相应的含义，明确这个变量的含义并给出变量，尤其是我们在网上抓取的这样一些数据。其次是识别关键的字或者关键变量，有着关键变量来把不同的数据集进行合并。

发言人 01:23:08

比如说我们在什么大家当一个什么超市老板，你这个连锁超市，大规模的超市，你可能有不同的数据对吧？有进货单，有销售的信息，有一些什么顾客会员的信息，还有一些财务的信息。这样一些不同的表单，你怎么样把它合并成一起。比如说你想要看这个月销售增长最大的这个商品类型是什么？那这个时候你怎么样来分析是不是要用到什么销量的信息，用户的信息。你想看什么样的人在这样一个这样这样一个销售，其他的什么数据购买情况。比如说有一个数据集里面，它就有一个什么consumer ID，他就可以和什么另一个数据集的这个customer number，它指的可能都是什么客户的编码。那这个时候我们就需要把这个密码不同的，因为它指代相同信息，但是用不同变量我们要把它进行一个统一。

发言人 01:24:26

其次我们要明确这个变量它的表示规则。比如说一些变量它的什么数据的形式，尤其是什么日期型的这样一个数据的处理。他有的时候比如说是年月日杠的形式，有的时候什么日月年这样一些形式。如果说我们不提前把它一些规则进行限定或者统一的话，你在进行数据分析的时候，他可能会识别错误对吧？他比如说会把这个什么一二杠、20杠、2020可能会识别成一二年对吧？他可能会造成一个最后分析的这样一个问题。

发言人 01:25:10

还有一些什么？比如说他数据里面可能会有一些缺失值或者空值，那这个空值或者说空白值，它它的这个什么应当怎么处理，或者说是当存在一些问号值、斜杠值、NA值、NULL值，那这些值你怎么样来对它进行一个进一个相应的处理。这也是在这个数据预处理部分需要注意的内容。

发言人 01:25:43

其次还需要去买避免这些冗余的信息。比如说一个数据集它有一个收入的信息income，那你合并完之后，它可能另一个数据集里面有个income com那这个时候它都是什么的可支配收入的信息。在数据合并的时候，你需要辨别投放，只保留其中一个就可以了。

发言人 01:26:10

大多时候就是冗余的数据，它是可以对这个什么相关检验可以检验到。比如说如果两个变量做这个相关分析的时候，它的这个相关性为一，那其实它就是完全质的两个变量值，对吧？那这个时候就可以去把它一个其中一个进行删除。此外如果是同一调查数据的多个数据集，或者是包含重复值的这样一个原始数据，或者说数据集合并之后产生的重复值，都要把这些重复值去识别出来，然后去删除。

发言人 01:26:50

此外还有一些数据值，它存在一些冲突，我们也需要去对它进行一个相应的这样一个检测与福利。比如说有一些收入，它在表里面他的这个什么计量单位是什么元，那在另一些表里面它可能是万元，那这个时候你就需要去把把这些亮光去统一，对吧？那在进行数据合并的时候，就需要把什么因为尺码编码这些不同，它的冲突你要把它识别出来。那除了我们这个什么数据值的合并，另一个比较关键的处理就是我们这个缺失值的处理。

发言人 01:27:38

缺失值在处理之前，我们就需要去判断这个缺失值它产生的原因。那一个产生的原因它是机器原因，另一个是人为的原因。机器的原因它可能是因为什么数据存储的失败。或者说是啊比如说你数据爬取的时候，刚好你爬取到这个网页这个指标的时候，他这个网络卡了一下，就没有爬下来信息。但是他也很快去检索，爬取下下一个点击的这个信息。那这个时候你可能就需要去算，把这个数据产生趋势值的原因去给它识别了出来。

发言人 01:28:24

人为的原因很多时候是因为什么你在做调查问卷的时候，有的它相应的问题。有的人可能鉴于隐私，或者说是一些别的考虑，或者说是这个数据录入人员他遗漏的这样一个数据的数据的录取，也有可能会造成一个人为的这样一个缺失值。那缺失值按类型来分，它存在一个随机的缺失和非随机的缺失。大家可以想象到随机的缺失的话可能是一个比较好处理的问题，对吧？

发言人 01:29:02

或者说可以这个什么还是一个什么，与其他变量取值无关。他这个什么随机缺失了，他是指什么？跟其他变量或者说没有缺失值的变量取值有关。比如说某项网络调查，他发现什么年龄大的，我发现很少有年龄大的受访者参与。那非随机缺失它是什么？不仅跟其他变量有关，它还跟什么它自身的取值也有关。比如说在收入调查的时候，这个收入趋势值可能是因为啥这个回答者他不知道如何量化自己的收入，也有可能是什么他的这个收入太高或者太低，他不愿意回答。那这个时候它就是一款非随机的这样一个缺失。

发言人 01:30:00

大家可以想象到，如果我在调查这样子收入消费的情况，当这个收入过高或者过低的人都不愿意去回答，或者说都不愿意收让我收集到这个数据的，它会存在一个什么问题？就是我的样本可能就没办法很好的去代表总体了，对吧？也就是我的最后的分析结果，或者我得到这个Y等于AXA加BX这样一个模型，它只能刻画出来收入在中等的这些人，他的这个反消费行为表现，就收入过高或者过低的。因为我这个数据缺失的问题，我都没办法多少。

发言人 01:30:43

那缺失值它也有一些相应的处理办法。最简单就是我这个数据空了没有，那我就把它删掉，对吧？那这个是最简单，是最有效的这样一个方法。

发言人 01:30:57

除了直接删掉，我们也有一些其他的办法去进行弥补。如果说你的这个缺失值非常重要，如果你把这一条信息删除，或者说这一些信息删除的话，对我们结果可能产生比较大的影响。那退而求其次，我们可以采取一个相对来讲稍微麻烦一点的方法是什么？我们可以去把这个缺失值来用我们估计的方法去把它补上，对吧？

发言人 01:31:24

假设说我们现在缺失，比如说我们拿到的数据？从2000年到现在，西电学生平均的这个收入情况。假如说我们缺失了2020年的这样一个数据值的话，大家可以想象到我是不是可以用一些数据去把这个值给，为什么？因为可能它会存在相应的趋势，对吧？那我可以把这拿什么2019年和2021年的平均值来作为一个代替，对吧？因为可能大家的这个收入可能会有一个比较比较什么合理的预期的这样一个趋势。那在趋势的插补的这样一个方法当中，除了均值，还有一些别的值我们可以用来擦除。比如说众数、中位数、均值都是可以用来。

发言人 01:32:23

除了简单的这样一个什么，我们小学学过的这样一个数值来计算得到进行差不多我们也可以用一些什么。极大自然估计热卡插补，或者说最近距离插补、回归插补、多值插补的这样一些不同的估计的方式，来把这个值估计值给它计算出来。如果说我们这个缺失值是数值的变量，那很多时候或者是最普遍的方式，我们是选择侦查平均值用来插补，也可以用什么平均值修正，比如用前后两个观测值进行插补。如果说缺失值不是这个数值型的变量，那我们可以选择什么重数进行插入。比如说我们现在调查这个西电收费情况，我选了这什么600个雪山的进行调研。最后发现其中有两个人我可能其他信息都有，就没有他的这个收入信息。那这个时候我一个合理的方式我把600个学生里面这个收入最出现次数最多的，比如说2000 1500块钱出现的次数是最多的那我就把这两个同学的这个码收入甚至可以用1500来替代。来补齐这样一个相应的缺失值。

发言人 01:34:02

同样我们也可以用极大自然估计。极大自然估计是我们通过观测数据的编辑分布，这个时候我们可以对未知的参数进行极大数量估计。这个时候，我们用这样一种方法，适用于一个大样本的情况。这样的方法它也有个缺点，可能会陷入局部的及时，而且收敛速度不快，计算也比较少。那除了这个，其他事人出具，也有可能有差最近距离，比如说我们用T减1的时刻，还差补T时刻的缺失，那这种插值的方法它不适用于时间序列中波动比较大的数据。

发言人 01:34:46

比如说我们现在收集大家的什么收入消费数据是按月的，你每个月都可报一下你的啥收入和消费的信息。比如说六月份的数据我有，但是七月份的我没有，八月份的我也有。那这个时候我是不是可以拿这什么六月份的数据来替掉七月份我缺失的这个数据值。

发言人 01:35:09

这个是最近距离的插补，还有什么热卡插补的方法，热卡卡布是给我们是不是有其他的信息？比如说我A同学他的什么，我没有他的收入信息，但是我有这个同学其他所有的信息，比如说他班级、他的年龄、他的出生地，他的所有其他其他信息。这个时候我可以在我600个学生的数据集当中，找一个跟A同学最相似的。因为他们来自相同的地方，相同的性别，然后相同的年级，就类似其他信息都相同的情况下，给我把他的这个收入情况用来把把它的趋势值进行一个插补的方式。

发言人 01:36:00

还有我们可以通过这个回归差别，也就是说我构建一个什么相应的回归方程。假设这个Y变量存在缺失值，但是我们知道X和Y轴之间的相关关系，这个时候我们就可以用回归的方法把这个码缺失值预测出来，然后对它进行一个填补。还有我们可以通过这样多值查，比如说我们它来自于什么？贝叶斯估计，就通常我们认为马岱插补的这个值，它是随机的，来源于什么？你观测到的值，那这个时候我们可以估计出相应的差异值，再加上噪声的几个一个随机的音效。这个时候可以形成多种可选的差不值，根据某种选择依据来选取合适的差不值。还有我们可以对它进行一个什么异常值的处理。数据质量它存在不准确和不一致的问题，往往表现出这个异常值。

发言人 01:37:13

比如说存在一些离群值，离群点上生。那异常值的检测基本思想是什么？我们从这个给定的数据集当中发现跟其他数据相比也有显著不同的数据。

发言人 01:37:27

异常值检测大概基于什么统计分布的方法，或者说临近性聚类的这样一些方法。异常值的处理也是首先我们可以基于统计分布的这样一个方法。基于统计方后边的话我们是不是学过什么我们的分布里面会有一些什么小概率的事件，比如说我们大家的什么收入消费情况里面，假设说大家这个消费信息平均值是吗？1500，然后方差是假设说是20，或者说是你是不是就可以通过这个码341个码法则，或者说是你可以求出这个分布5%的这样一个小概率区间，来算出这个数据大于加减三个标准差的这样一个内容。比如说你平均值1500，然后你这个标准差是20的话。那假设说你现在数据集里面有一个同学，他的什么消费一个月是六千。那这个时候你就可以可以识别出来这个数据可能是啊有问题，对吧？可能是一个异常值，这个时候你就可以需要对这个异常值进行相应的处理。

发言人 01:38:52

还有有一些什么临近性的方法。临近性的方法它可能假定这个数据对象它是一个异常值。如果说他在特征空间中最近这样也远离他，你说这个对象与它最邻近的邻近性也显著的偏离这个数据集中的其他对象，他就认为这样一个数据是一个异常值。也有我们可以基于聚类的这样一个方法，就假定正常数据对象它都是在这样一个大的范围内比较稠密的这样一个促。那异常值它可能就小了，吸收的醋或者说他自己一个人成为这一个点，成为了一个数。

发言人 01:39:39

除了数据异常值，我们也有一些数据变换的这样一个处理。在数据预处理这样一个阶段，我们需要把什么数据往往进行变换的操作，去消除数据之间亮光的问题。然后把这个数据变换到这个正态分布，让它更加的规整。这样我们建模得到的结果才会更加的准确。那对这个数据进行变换处理有一些什么？比如说数据光环、变量构造，数据规范化、离散化，比如说数据概化这样一些方式，把它整合成一个什么正态分布的这样一个手册。我们这节课先上到这。

发言人 01:41:48

我需要交送到谁家。