

订阅DeepL Pro以编辑此演示文稿。  
访问[www.DeepL.com/pro](https://www.deepl.com/pro?cta=edit-document)，了解更多信息。

好的大家早上好我的录音开始了这意味着我也得开始了十点准时开始我知道是的，自动开始所以所有的讲座 录音都会在预定时间自动开始 所以你必须在那之前准备好否则你就得开始剪切录音了而这只是一个整体的其他。我知道我不知道为什么要这么做这是我工作中很奇怪的一部分好吧 谢谢关心 Okay.谢谢你的照顾大家都好吗？还不错差不多吧爱丁堡的天色开始变暗了 所以不习惯阴暗潮湿的冬天的朋友们做好准备马上就要开始了是啊，现在已经十月了。所以到了十一月，你们都会怀念八九月份的好天气，那时候的苏格兰比你们平时想象的要干燥得多。所以我希望你们已经准备好了 So I hope you're you're ready for that.尽量多花些时间在户外。如果需要的话，还可以买一盏日光灯。显然，它们真的很不错。是啊冬天尽量保持乐观如果你跟医生说过，可以补充维生素D，因为这也有帮助。那就好希望你一切都好本周我想做的第一件事就是简单谈谈你的课业。我知道这似乎是一个非常非常长的路要走，因为这个课业的截止日期是 12 月初，但我还是想开始谈论它，因为我认为，如果你现在开始在你的小组内考虑它，那么最终会更容易及时地赶上截止日期嘛。你们要记住，下个月你们还要完成数据分析原理的最后期限，所以在那段时间里，你们不会有太多时间来做这个课业，这很正常。所以你必须同时管理这两个项目。如果你及早开始，这是可以做到的，而且你也知道，我想几周前我们已经上传了一份简报。我还上传了数据集。现在有两个文件。我不知道你们是否有机会看一看。主要数据文件是 TXT 格式的信用评分数据，然后是变量列表，这是一个数据描述文件。如果你已经看过了，你会发现它并不漂亮。它看起来并不好。你可能对我很不满意。这里有非常奇怪的缺失数据、非结构化数据、奇怪的变量描述，这些都是不完整的，而且不是很容易理解。这是故意的。所以，如果你觉得我犯了错，那是我的错。我也是人。我在为你创建数据集时确实会犯错。但如果你觉得这很奇怪，那很可能是故意的。所以，如果你看看你的变量列表，你会发现每个文件都有这些变量描述。让我们在 Excel 中打开它。这样就容易多了。嗯。真的吗？好的好的，这实际上是一个非常典型的数据描述文件。你可以看到，B 列是变量名，C 列是描述，D 列是变量值。所以这都是故意的。举例来说，如果婚姻状况编码没有告知，而你又知道它是某种序数变量，这就意味着你对该变量的数据描述不完整。因此，在这种情况下，我希望你们能做出合理的假设，并在这种情况下进行一些批判性思考。例如，你可以假设这些类别可能意味着什么？或者你可以说，我无法对此做出合理的假设。最好忽略这个变量。或者你可以说，我们先做一个分析，然后他们在解读结果时，可以自己对结果进行解读。因此，你可以有很多不同的方法，但这样做是有目的的。所以，如果你觉得这很糟糕，这很奇怪，那就这样吧，你得好好处理。现在，如果数据中真的有错误，你会想，好吧，这不可能，它完全破坏了我的代码，或者你甚至无法打开数据集或类似的东西，请让我知道，因为数据集中可能真的有错误，但如果它只是有点不方便，那么它是故意的。好的有没有关于数据的初步问题？有没有人看了数据后已经绝望了？没有有人点头我看到有些人已经看过了。你们觉得怎么样？好数据数据不好可以的是的，这并不容易，但我认为它可以帮助你很好地理解现实世界中的数据是什么样的。所以，这门课程的目的就是给你处理现实生活中的数据所需的工具，这些数据可能看起来并不漂亮。另一件事是，你可能会想，我还不能开始做任何事情，因为我们在讲座中还没有做过任何建模。我该怎么做呢？但我认为，如果你回想一下前两堂课，每堂课实际上都涵盖了一些非常重要的内容，这些内容将成为报告的重要组成部分。因此，我认为你完全可以开始着手撰写报告。例如，回想第一讲，我们在讨论如何提出研究问题。什么是研究问题？对于什么样的研究问题，我应该使用什么样的模型？你可以开始描述数据可能来自的背景。在开始对模型本身做出决定之前，模型甚至可以告诉我什么？你必须先决定什么样的模型适合这个数据，或者我想回答什么样的问题，这些都来自于简介。所以，你已经可以开始回想第一讲的内容了，也可以开始回想第二讲的内容了，在第二讲中我们谈到了缺失值、数据估算和异常值等等。正如我们刚刚注意到的，数据中可能有很多这样的内容。因此，你也可以从这些预处理步骤开始。所以，如果你很想尽早开始，你可以这样做。话虽如此，本周我们将开始讨论几个你通常会采取的非常简单的预测建模步骤。本周我们将讨论线性回归。这是理解预测模型如何工作的基础。我们之所以把线性回归放在最开始，一是因为它与你在其他讲座中的内容、你过去所做的事情有着自然的联系。另一个原因是，用线性回归来解释预测模型如何工作真的非常非常容易，因为这是一个非常直观的研究模型。好的在开始之前，我们先简单回顾一下上周的内容。正如我提到的，我们讨论了数据清理。我们谈到了数据集的准备、预处理标准化、变量转换。你还记得我们在做离群点检测时，对变量进行的整数热编码吗？这个想法就是，看看我的观测值和缺失值的平均值与平均值相差多少标准差，我们如何才能去除它们或如何才能估算它们。本周，我们将首先对描述性统计和可视化做一个简单的介绍。我稍后会解释这一点，然后我们将用简单明了的步骤，深入浅出地介绍线性回归。这将与第五周的 "数据分析原理 "讲座同时进行，届时我们也将学习线性回归。但在该讲座中，我们将更多地研究该方法的假设和统计基本原理。而在这里，我们将更多地探讨该方法的适用性，以及在预测性设置中使用该方法时，比在描述性设置中使用该方法更需要记住的事项。因此，我们将讨论简单线性回归、多元线性回归以及你必须考虑的任何应用和注意事项。是的，我说了两遍。那么，我们为什么要讨论描述性统计和可视化呢？同样，我们在每一堂课上都讲过了。我认为到目前为止都是如此。原因在于，在任何一种数据分析、模块、项目、研究、问题中，描述你的数据都是非常重要的原因。其中有些我们已经谈过了。因此，我们谈到了描述性统计、描述性统计的可视化如何用于数据的可视化和交流。我们已经强调了这一点有多么重要，因为它可以帮助你弥合你的技术知识、对数据和建模的技术理解与你试图与之交流的人的理解之间的差距。因此，如果你想与某个人沟通，比如管理层，那么你可能要想一想他们想回答什么问题，并尽可能简短、快速地找出答案。他们不会有。如果你必须向更高层的人汇报，他们很可能没有时间阅读 20 页的报告，尤其是如果报告中有大量的数字和关于模型和所有这些事情的大量细节。他们可能会问你，好吧，你想让我从这份报告中得到什么重要启示？如果你指着一张漂亮的图表，解释说就是这张图，他们就会对你非常满意。所以，每次做任何分析或撰写任何类型的报告时，都要试着想想你的受众。你的受众是谁？你想向谁报告？回想一下这次的作业评估 在这门课程中，你的受众将是一家银行。那么，想想银行到底对什么感兴趣？他们对模型的大量技术细节感兴趣吗？他们可能会感兴趣，至少你已经足够谨慎地进行了模型设计，并对假设进行了测试，等等等等。他们会对模型的有效性感兴趣。但回想一下，你实际上并不想把答案和结果展示给那些对统计背景不感兴趣的人。这就是可视化交流。这是一个很大的课题，尤其是在商学院的环境中。另一件事是检测异常值和数据结构。我们刚才谈到了离群值，离群值基本上是指分布的两端。但你可能还想检测出只出现在数据集特定部分的异常值。现在，这个特定部分可以是分布的尾端，也可以是你从不同来源收集的数据，而某些来源的数据缺失程度较高。值。理解这一点对你来说非常重要，因为想象一下，你从两个不同的来源收集数据，而你认为例如这两个地点对理解数据的整体结构同样重要。举例来说，如果其中一个数据集的某一列有全部缺失值，而你因为了解到缺失值而删除了所有行，那么我们就删除它吧，因为这一列很重要。你基本上是完全删除了从其中一个地点收集到的数据，这会使你的模型偏向于只对其中一个地点做出响应。因此，要了解异常值的位置，既要了解它们来自哪些数据，也要了解它们在哪个变量或哪个区域的数据中排队，例如，如果你有一个数据集，你正在查看一个人的收入，你注意到高收入组中的每个人都没有报告他们的教育程度，那么这也会使你的模型产生偏差。如果您对教育这一变量感兴趣，那么突然之间，您就只能分析样本中低收入人群的教育分布情况了。因此，请思考异常值出现在哪里，然后思考它们出现在那里的原因。我们还必须了解数据可能的分布情况。我们在校长讲座中已经开始讨论这个问题。因此，出于多种原因，思考数据分布、数据来源是非常重要的。检测异常值是其中之一，为模型做假设是另一个原因。我们还将讨论如何利用这些假设来评估模型性能，并了解数据的总体趋势和相关性。因此，这正是我们今天再次讨论可视化和描述性统计这一主题的原因。我们将看看如何使用可视化来评估模型性能。我们还将看看如何利用可视化来检测数据的总体趋势和相关性。例如，判断我们的数据是否是线性的。如果你回想一下我们今天研究的主要模型--线性回归，你就会明白判断数据是否真的遵循线性趋势是多么重要。让我们从评估模型性能开始。这是系列讲座中会反复出现的主题之一。因此，每次讨论一个新模型时，我们都会简短地谈谈如何评估模型。原因在于，例如，根据模型的实际预测效果，有不同的方法。因此，根据不同的方法、不同的模型，我们可以使用非常不同的方法和非常不同的衡量标准。今天我们将介绍两种方法，因为它们常用于线性回归，你也会看到它们反复出现，因为它们是非常重要的概念。第一个是均方误差 MSE 以及其平方根，即均方根误差 MSE。这基本上是一种误差率。因此，我们要看的是任何预测值和任何观测值之间的偏差。如果我说预测值，我指的是一种松散的意义。因此，今天，我们在讲座中使用线性回归的方式并不一定是作为预测模型，就像我们试图非常具体地预测新数据集带来的新记录一样。但我的意思是，我们将看到如何为该模型导出参数，从而预测不一定相同的新数据记录。你会看到区别的。你会看到区别的。我所说的线性回归预测和新观测数据预测，就像我们今天所说的那样，下周我们将讨论逻辑回归。所以，请记住这一点。一般来说，我们只需查看预测值与观察值之间的偏差，并以此来评估模型的整体性能。整体是这里的关键词。另外，我们还要看 R 平方，它使用解释方差的大小来评估模型的整体性能。稍后我们将对此进行更详细的介绍，但首先我们要看的是均方误差。正如我刚才所说，MSE 是衡量整体误差和模型性能的指标。因此，如果模型的结果是数值型的，你就可以使用它。记住这一点非常重要。原因在于，我们基本上是在计算观察值 y 与预测结果 y hat 之间的距离。然后，我们把所有这些偏差、所有这些小误差加起来，再除以观测数据的总数。然后取其平方根。因此，如果取其根，如果取 MSE 值，就能得到与样本中使用的单位相同的偏差。因此，在很多不同的软件输出中，你都会看到 MSE 或其平方根。因此，这是一个相当重要的数值。现在，希望你能马上看到我说了多少次这是总体，这是总体。记住这一点很重要，因为它并不能告诉你模型在特定数据部分的表现如何。因此，我们将在整个讲座中讨论不同类型的误差。因此，重要的是要记住，有时某些类型的误差比其他类型的误差更严重，这取决于你试图预测什么。举例来说，如果你试图预测病人的病情，但却漏掉了某个人，这就比预测某人可能患病，但后来发现他并没有患病更糟糕。因此，不同程度的误差比其他误差更容易被接受。在这种情况下，我们只是从数字意义上看所有的偏差，是整个数据集的总和。我们并不关心某些区域的误差是否大于其他区域。我们只是将它们全部发送出去。我们将它们除以记录数。我们基本上查看整个数据集的平均误差。那么实际情况是怎样的呢？如果我们想一想，我们可能会有某种散点图，例如，我们的 X 和某种值 Y 之间的散点图。真漂亮比方说，我们预测 X 和 Y 之间存在某种线性关系，然后我们要做的就是把所有这些偏差相加，用不同的颜色表示出来。那就太酷了嘿，又来了所有这些小偏差形成实际预测值。因此，如果将所有偏差相加，再除以记录数，就得出了均方误差。我们的想法是的。你已经明白我想说什么了。如果我说你只看整个数据，你就会明白我在说什么。那么让我们想象一下，让我们换一种颜色。我喜欢紫色。我喜欢紫色。让我们想象一下，其中的一些，尤其是这里，比这里差很多所以你可以看到上面的偏差 比下面的偏差要大得多 而下面的偏差非常非常小但如果你从整体上衡量，你不会发现这一点。在更高的空间里，你也不会发现这一点。偏差实际上要大得多。你唯一能发现的是，你的预测值的平均偏差基本上是一个数字。因此，它们可以用来衡量模型的整体性能。但由于我们是对所有数据集的所有残差进行求和，因此我们实际上并不是在研究其特定空间的偏差。因此，我们并不知道模型是否在某些方面表现良好，而在其他方面表现较差。而实际检测模型是否在某些空间表现良好而在其他空间表现较差的方法之一，就是将其可视化。所以，我们刚才看了一下散点图。例如，你可以看到数值越大，偏差就越大。如果你将其可视化，你就可以看到这些偏差越来越大，然后你就可以说，嘿，我的模型在 X 值较高的区域表现不佳。所以，如果你的数据正在扩散。那么在较高的数值中就会产生更多的变化。而你仍然有一个预测值。这只有通过可视化才能看到。如果只是计算误差，则无法看到。我认为这也是我有时在机器学习研究领域遇到的问题之一。我认为人们在使用机器学习模型时，往往会过度强调误差率的重要性。强调误差率和准确率对数值的重要性。只要看这一个数字，基本上就能告诉我这个模型整体表现良好，而这一个数字就能证明它绝对是正确的模型。因此，他们可能会选择一个模型，而不是另一个，例如，因为它的准确率更高，而不会考虑它是否在数据集的所有区域都表现良好。如果数据集的所有区域都同样重要，那么他们就会选择这个模型。举例来说，如果你有一个模型的总体误差稍差，平均平方误差稍高。因此，误差率较高，模型较差，但在整个数据集上表现同样出色。那么在某些情况下，即使它的误差较小，也会是更好的选择，因为它实际上会同时强调这两个空间，比如这里。因此，如果你有一个模型能够同时捕捉到这些区域和上面这些区域，那么尽管它的误差率最差，也可以说它是一个更好的数据模型。所以，这也是我最讨厌的一点，就是简单地报告一个单一的错误率或单一的 R 方值或准确度值，不管你报告的是什么，而不考虑是否对所有记录都适用。所以，是的，小毛病。请记住这一点，也许以后你自己建模时会用到。哦，天哪。哦，不我在说话今天的教训我不能呼吸咖啡 即使我想。哦 Oh.如果你们都能呼吸到咖啡因 那岂不是太棒了？哦 好吧 Oh, okay.说回这个我想说的另一个指标是 R 方。它可能是最著名、最重要的模型性能指标之一。你几乎随处可见它的身影。如果你想在 R 中建立任何模型，都会得到某种 R 方。如果你阅读任何一篇关于建模的论文，他们很可能会报告一个 R 平方值。因此，它几乎无处不在。能够解释它是一项非常非常重要的技能。所以，你会看到关键词。同样，这是一个整体模型性能指标。因此，我们不查看数据的特定区域。我们创建了一个整体模型性能指标，计算方法与之前的误差率非常相似。因此，我们仍然需要将观察结果 y 与我们对特定值 I 的预测结果的偏差相加，然后将其平方，等等。不过，现在我们还要用观测值除以该数据的平均值。然后，我们用 1 减去这里的整数。而另一种说法实际上就是这里最下面的那个。因此，你会多次看到这种方式或类似的方式，这取决于你如何实际拼写误差平方和。因此，R 平方也可以描述为平方误差之和，也称为残差平方和和总平方和，即总平方和。因此，你还会看到 SS 和 RSS，但这是同一个概念。相除。然后再减去这个值。基本上是一样的。所以我们刚才说的 SSE 是我们的平方误差。因此，所有的偏差，例如，从我们的回归线，然后我们的 ssts 是平方总和。这基本上意味着它是我们对平均值的偏差。那么，数据围绕平均值的偏差有多大？我们在这里可以看到两种不同类型的变化。其一，我们看预测值周围的变化。例如，我们的预测线，我们的点分布有多远？另一种是总体变化。那么，如果是平均值的话，这些值的分布范围有多大？不，还没有。显然还没有我还没画呢我为什么不画？真奇怪嗯 Hmm.那你只能用我的手来描述概念了 So you'll have to do with my hands describing concepts.好吧比如说，我们有线性预测因子周围的偏差，我们有总体偏差，总体偏离平均值的偏差。通常以模型解释的变异量占数据总体变异量的百分比的形式进行报告。因此，如果你在文献中看到 R 平方的报告，在很多情况下，要么报告为 0.7，要么报告为 70%。然后论文可能会说，该模型解释了数据总变化的 70%。因此，我们有一个总的变化，围绕平均值的变化。我们究竟能解释多少周围的变化？例如，线性预测器。它和我们的 MSE 存在同样的问题，因为它是一个整体评估工具。此外，对于预测因子较多的模型，R-square 的准确性也较低。因此在这种情况下，通常建议使用调整后的 R 平方来代替。计算公式在这里。如果你有更多的预测因子，它们自然就能解释更多的数据变化。因此，预测建模中实际引入的数据越多，自然就能解释更多的变化，尽管差异可能非常微小，尽管引入更多的预测因子有时也会导致数据过度拟合和类似的问题。因此，调整后的 R 平方基本上考虑到了预测因子越多不一定对模型越好，因此会对预测因子的数量进行调整。因此，如果您有大量的预测因子，请选择调整后的 R 平方。如果预测因子较少，也无妨。因此，并不是说只有超过 30 个变量或类似变量才能使用调整 R 平方。如果你的 N 和 UK 之间的差异基本上很大，或者差异不大，那么你可以用任何方式来使用它。所以，无论你是否有大量的预测因子，这都是一个安全的选择，但如果你确实有大量的预测因子，你就必须使用它。好了，让我们来看可视化重要性的第二部分。我们谈到了选择模型在很大程度上取决于对数据的理解。这意味着多种不同的东西。但在这里，我指的是数据的分布，更具体地说，是数据与你想要预测的结果之间的关系。因此，我们之前从偏离线性预测结果的角度来看了 MSE，基本上就是偏离我们的直线。那么，我们如何知道我们应该使用线性预测因子呢？有时，最好的方法就是绘制数据图。因此，第一步绘制数据图确实有助于你了解自己是想使用线性预测器，还是必须使用非线性预测器。举例来说，如果你有散点图看起来是这样的。你还记得我们在第一讲中谈到聚类的时候吗？我们谈到，嘿，数据中可以看到一些线性趋势。所以，这就是我们能看到一点线性趋势的案例之一。就是这样。所以在这种情况下，我会说，好吧，试试线性模型。它可能会给你一个很好的拟合。在其他情况下，如果你的关系看起来像这样。嗯嗯嗯嗯嗯嗯。我在学校专注于艺术。所以我是一个糟糕的艺术生。现在你会发现这对我的工作很有帮助是不是很棒？所以这实际上是一种非常非线性的关系。在这种情况下，你需要寻找一个非线性预测模型，比如决策树，它通常效果很好。我们再来看看第三种情况。我很好奇你会怎么做。让我们先把这个去掉一点。给我一点空间坐吧是的，是的，是的，好的如果你有这样一个模型，你会怎么做？嗯？让我做一下嗯，不错你觉得怎么样？像这样？像这样？是啊，你绝对可以适合这样的东西。所以，你会适合， 像，非线性的东西。基本上就是你说的那样像这样？很好 - Yeah.非常好还有吗？另一个想法好吧，如果你把这些分开来看，它们看起来确实非常非常线性，不是吗？这里是线性关系，这里是线性关系。实际上，你可以建立一个片断线性模型。所以你可以说，你在这个值空间拟合了一个线性模型，然后又在这个值空间拟合了另一个线性模型。因此，有一些使用分段线性回归的方法，例如，它可以让你非常接近地拟合在你的价值范围的不同空间中具有线性关系的数据。为什么要这么做？有谁能想出一个好理由吗？便宜。是的，这真的很简单。它简化了事情，这意味着它更便宜。因此，例如这种回归，然后对数据的不同空间进行拟合回归，在建模方面会更便宜，尤其是当你有一个庞大的数据集时。另一个优势是，我们会发现线性回归可以很好地进行解释。因此，如果解释是一件大事，那么考虑如何将小的线性回归拟合到数据中可能会有所帮助。如果你有兴趣，还可以查查另一个东西，那就是样条曲线。这真的很有趣。我想有人告诉过我，你的一位同事之前问过我关于时间序列分析的问题。比如说，你有一个时间序列。会发生什么我不知道我不知道时间序列是什么样的，诸如此类。实际上有一种方法，你可以尝试将样条曲线拟合到数据的不同区域，所以你会拟合它们，可能是类似的东西，然后将它们拼凑在一起。因此，这是与时间序列数据相关的另一个非常有趣的事情。这也是一个类似的思维过程，因为你将这些片段拟合在一起，以复制，在这种情况下，时间序列是如何发展的，而在底层情况下，数据是如何具有线性形状和数据空间的不同区域的。因此，线性线性模型，并不总是你要寻找的这种非常简化的直线关系。它可以拼凑得非常整齐。简要提及偏差方差。在这种情况下，我们将在下周进行更详细的讨论。不过，开始思考这个问题可能相当重要。因此，高模型偏差可以被认为是模型过度简化了某种关系。因此，这基本上意味着，到目前为止，即使添加更多数据，模型也不会改变。这是一种非常简化的关系，而高模型方差可以认为是模型将我们的关系过于复杂化了。因此，如果在此基础上增加一些小的数据点，模型就会发生很大的变化。换个角度思考，如果我认为模型偏差大，我就会告诉自己，它很稳健，但不灵活。如果你认为模型方差大，我认为它很灵活，但很敏感。因此，对我们来说，关键在于找到一个能在两者之间进行最佳权衡的模型。因此，我们需要的是一个相当稳健的模型，这样即使我们增加一两个额外的数据点，它也不会改变，但我们仍然希望它能与我们的数据拟合得相当好。这与我们之前的图片有什么关系？如果我们忘掉样条线的概念，只是认为这是我们的回归线或建模线，你可以看到这条线是如何紧跟我们的数据，以及这条线是如何精确地紧跟数据的。现在，如果你稍后进行修改或有额外的数据点，那么这条线就会发生很大的变化，因为它是如此紧贴你的数据点。而在这里，如果你在这里或那里增加几个点，并不会真正改变这条线。直线就这样穿过它。因此，这种数据，例如左上角的数据，我们称之为模型偏差较大的模型。所以它没有明显变化。如果我们添加任何东西，它都不会非常接近数据。我的意思是，它相当不错。这是一个非常好的模型。虽然是我画的，但如果我们添加更多数据，它的形状可能不会有太大变化。因此，我们会说它不是很灵活，但它是完整的半身像，这可能是有用的，这取决于你的上下文。因此，请牢记这个词--偏差 变异 权衡 在整个系列讲座中，你会多次看到这个词。例如，它还与过度拟合模型的概念有关。因此，如果你过度拟合一个预测模型，这意味着我们将它与我们所拥有的数据（即训练数据）拟合得非常非常紧密。但是，如果我们尝试将其用于未见过的新数据，并进行微小改动，那么模型就会崩溃，因为它过于紧紧抓住数据应该是什么样子的理想化想法。好了，现在我有个问题要问你。变量的形状对建立预测模型有什么重要意义？所以 A，你需要知道因变量的分布形状。B 你需要知道自变量的分布形状，才能建立模型 C，你需要知道自变量和因变量的分布形状，才能建立模型 D，你需要了解因变量的分布形状，因为是否需要知道因变量的分布形状取决于你选择的算法。你认为哪种算法最准确？你是说 B 还是 D、D、D？谁想回答？好的好吧谁认为答案更好？没有，还有其他意见吗？有所以可能是你我认为这是最准确的。我认为这是最准确的所以，如果你确实需要了解该模型的数据形状，那么你就需要了解两者的形状。如果你选择的模型认为变量分布的形状根本不重要，那么你可能更接近答案 D。它们是一个相当好的例子，我认为数据的分布并不重要。例如，我们不必检查线性度。不过，我们也了解到，变量的分布也能告诉我们异常值的情况。因此，你可能会说，如果你使用变量的分布来检测异常值，那么即使对决策树来说，它仍然很重要。决策树对异常值并不是超级敏感，但如果决策树不试图对存在异常值的特定变量进行拆分，它们就会变得特别敏感。因此，在这种情况下，即使不考虑模型本身的假设，你也需要了解变量的分布情况。好了，现在是 10:45。我们现在休息十分钟，因为这样我就可以把所有的回归内容放到下半部分，这样在结构上就比较合理了。10分钟后见好吧，我们要。让我们回来。请坐请坐请坐请坐 请坐 Sit down.Sit down.Just.就这样好吧，我意识到我没有停止录制，这意味着我以后必须回到录制中，剪掉中断部分，这很累人。我也是一名视频编辑。老实说，如果你在大学工作，你会同时兼任所有工作。这太疯狂了。我既要做研究，又要做很多行政工作，还要教书，显然我还要做视频编辑。没错，这就是我休息时的抱怨。另一件事是，有一个问题出现了好几次，是关于描述和统计可视化的。如果你有很多变量，你真的会去检查每个变量吗？它们与预测因子之间的关系如何？答案是视情况而定。这真的取决于你的时间。这取决于你要使用的模型。取决于这种关系对你的模型假设有多重要。还取决于你所拥有的变量数量。如果你要建立一个有 200 个变量的线性回归模型，除非你有时间，否则你不会去检查每个变量与预测因子之间的关系。之所以要这么做，是因为正如你将看到的那样，模型本身的结构假定预测变量与结果变量之间存在线性关系。因此，如果这种关系并不存在，因为它是线性的，那么拟合线性回归就不会给你带来最好的预测结果。尽管如此，如果我们实事求是的话，除非你有时间，否则你不一定会去检查每个变量的假设。因此，有时你可能会想，好吧，让我们试试线性回归，因为它非常快。例如，我们也可以在相同的数据上运行随机森林。让我们比较一下两者的结果。哦，看，随机森林的表现要好得多。原因可能是预测变量和结果之间不存在线性关系？有可能。然后你就可以去检查这是否真的是问题所在了。现在，我告诉你们这些，我能感觉到我的统计学家同事们正怒视着我，因为从技术上讲，线性回归有一系列假设，如果你违反了这些假设，那么你就不能使用线性回归模型，因为它不会给你一个有意义的输出。现在，因变量和自变量之间的线性关系并不是使用线性回归必须满足的严格假设之一。不过也有这样的假设。我们稍后会在幻灯片中看到。因此，从技术上讲，我告诉你们，因为我是你们的老师，我必须教你们东西，从技术上讲，在使用线性回归之前，你们必须检查我们将在这里看到的每个假设的统计有效性。现在我退一步，以一个在现实世界中经历过的人的身份告诉你，你应该检验你的假设的有效性。因为如果你不这样做，你就会开始陷入这样一种境地：让我们使用任何模型而不考虑假设，让我们使用模型并检查结果。如果你开始这样我不知道该如何形容。你就会陷入这样的陷阱：先是对一个回归模型这么做，然后对其他模型也这么做。那么你最终就无法真正使用你的结果，因为它们在统计上是无效的。因此，你得出的结果经不起推敲。所以，是的，你应该检查每个变量的假设，即使变量很多。但你并不一定要检查每一个变量的因变量和自变量之间的线性关系。如果你有 500 个变量或类似的东西。那么这对你的课程意味着什么呢？如果你使用的是线性回归法，那么谨慎的做法是，如果你的课文中没有列出所有的假设，那么就按照我们在这里列出的假设步骤来做，因为字数有限，我不会读超过字数的内容。然后，你可以在附录中说明你已经验证了这些假设。这里有图为证。这样你就可以向我证明，你知道，这些假设是很重要的。你检查了它们，因为这是一所大学，你必须经过四个多步骤来做分析。好吧，我希望这或多或少能回答你的问题。如果我有很多数据，我会在描述性统计中做一些事情，比如在正文中展示一部分数据，以此证明我做了分析。例如，我可能会选择一两个有趣的变量，展示一个散点图来证明，是的，我检查了变量的分布。他就是证据。然后我把所有其他的图都放到附录中。如果有人想检查我是否真的做了，我已经做了。好的说了这么多 我们来看看线性回归的理论吧让我们从简单的线性回归开始。我将这两堂课拆分开来的方式是，我们这周做线性回归，下周做逻辑回归，每次都做简单和多重版本的线性回归。最初的讲座结构实际上是让我们先做简单线性回归和逻辑回归，然后在下一次讲座中做多元线性回归和逻辑回归。所以请告诉我这样的结构是否更好，因为我认为今天只讲线性回归，下周只讲逻辑回归更有意义。不过，我也欢迎大家提出意见，看看是否将这两门课混在一起讲更有意义。线性回归。我希望你们都还熟悉这种方程。我们要做的是在二维空间中拟合一条线，这条线由一个单一的自变量和一个单一的因变量组成。我们试图拟合出一条最接近所有数据点的直线。因此，等式可以非常简单地表述为：我们试图预测 Y，即 AI 因变量的自变量值，我们试图用截距贝塔 0 来估计 Y。我们有 x i，即独立数据的第 i 个观测点，我们在这里有一个实际参数，它描述了 X 与 Y 的关系或输入对 Y 的影响有多大。然后我们还有一个误差项，因为人无完人，我们的模型也不完美。因此，正如我所说，线性回归拟合的目标是，我们试图估算出这两个参数，即我们的截距（beta 0）和更好的参数（描述与 X 关系的参数），以最小化数据点与直线之间的距离。大家应该还记得之前的情况，我们有一个二维空间，我们有数据的散点图，我们正在拟合一条直线，我们试图最小化偏差，最小化每个点与直线之间的误差。因此，这可能意味着我们要稍微移动一下直线。如果我们改变直线的角度，就是改变我们的贝塔系数 1，如果我们上下移动直线，就是我们的截距贝塔系数 0。因此，我们可以将这条估计回归线写成我们的贝塔，写成我们的 Y 帽子，也就是我们估计的 Y 值，我们利用更好的贝塔 0 帽子、估计截距和贝塔 1 帽子之间的关系来估计它们。我们的估计值就是我们预测变量的估计参数。哇我们还可以看到下面这个词。这就是我们的实际值 Y 和估计值 Y hat 之间的偏差。我们也称其为残差。它们基本上就是我们的误差。任何我们无法解释的关系，以及我们知道是真实的完美关系，但其中总会有一些误差。无论这个误差是什么，都是我们的估计值与真实值之间的偏差。好的，是的，你会记住的。我们还是把它画出来吧，因为这是一个很重要的概念，而且我很喜欢画画，因为我知道怎么画了。我仍然为自己感到骄傲。好了，那么。我们已经说过了。比如说，我这里有一个散点图。我们可以看到这是一个线性关系，所以我们通过它画出回归线。这是我们的截距。这就是我们的 Beta 值。它基本上描述了 X 为零时，数值的高低。所以，如果 X 为零，我们的值会有多高。哇，太可怕了好吧为什么还是零呢？在这种情况下，我们的默认值和 X 的斜率就是我们的更好值。所以，如果你仔细想想，这可能是一个非常平的或非常的英语单词，是什么呢？就是斜坡，与平坦相反。倾斜倾斜的所以是强烈倾斜这里有人会说母语吗？有的 - All right.是的，非常倾斜 Yes.非常倾斜的斜坡就是这个 That.That's what I mean.基本上，这就是一个高贝塔值。那么问题来了，X 对 Y 的关系或影响有多强？现在，这只是简单的线性回归，稍后我们将看到多元线性回归，我们将把它扩展到更多的维度。这可能是我画图的极限所在。我们究竟该如何找到这条最佳拟合线呢？我们刚刚决定，我们必须估算出贝塔零和贝塔一。你不能估计残差。那只是最后剩下的部分。因此，估算参数最常用的方法之一就是使用最小二乘法。因此，问题是在所有可能的数据拟合线中，例如通过散点图上的点，找出一条能使平方误差之和最小的线。这非常符合逻辑。我们之前想过，好吧，每个点与直线的偏差。如果我们把所有这些加起来，就是平方误差总和。所以，让我们把发现的误差最小化吧。大家应该还记得，我们之前讨论过只看平方误差之和的问题，即找到一条最符合数据整体的直线。因此，我们并没有考虑数据中的任何偏差和特定区域。我们只是试图拟合出整体上最好的直线。那么，用数学术语来说，这看起来像什么呢？我们看的是平方误差之和 E，可能应该是ε。好的。我们知道，这是每个真实值与每个预测值的偏差平方和，是所有数据点的总和。我们还知道，我们刚刚决定，我们的 y 帽子实际上最好用这个线性表达式来描述，即更好的 0 帽子加上 1 倍 x 的 beta 帽子。这就是我们的平方误差，由于平方误差最小，显然我们要尽量减小误差。那么，能使这个表达式最小化的贝塔 0 和贝塔 1 的值是多少呢？现在，我将向你展示如何精确地手工计算。请抓紧。如果我们利用问题的一阶条件，你应该还记得在学校的时候如何找到最小值，即利用一阶条件，将其设为零，然后求解变量。因此，在这种情况下，如果我们求解这个漂亮的表达式，就会有一个求解贝塔零的表达式和一个求解更好的表达式。如果你在纸上完成所有步骤，就会得出这样的条件：贝塔零的帽子应该是观测值的平均值。减去贝塔一乘以 x 平均值。因此，这些小横杠总是表示你的平均值。第二个条件告诉我们，贝塔一帽子可以最好地描述为每个值与其平均值的偏差。将每个结果变量与平均值的偏差除以每个值与平均值偏差的平方，然后求和。如果你满足了这两个条件，那么你就能找到更好的零和更好的一的最小平方估计值。所以，如果你需要手工求解线性回归，我不知道你是否需要。我从来没有手工做过。这就是实际操作的方法。所以有一种手工求解的方法。从技术上讲，也有手工求解多重线性回归的方法，但它们会变得相当复杂。所以在某些时候，你不会这么做。但还是有办法的。所以，如果有人告诉你，由于某种原因，你坐在一个没有电脑的孤岛上，但你想计算日落时间与椰子生长之间的关系，或者类似的东西，你实际上可以用手来做。我这里也有一些例子，我想可以很好地说明这一点。如果你能在右侧看到一个样本数据集，就会看到这样的效果。例如，这是一盒销售数据，可能是第一部电影的销售情况，以及对电影院盈利能力的影响。比方说，X 是我们的电影销售额，Y 是我们公司的盈利能力。你会发现它们都比较低。事实上，我已经做到了。好了，现在是故事时间。实际上，我在学生时代就在电影院工作过，因为我非常喜欢电影。我当时在一个小镇的电影院工作，实际上有两个地方。那是一个非常小的巴伐利亚小镇，我们有两家分店，每当我们人手有限的时候，我就得在两家分店之间来回跑，然后他们的巧克力棒用完了，我也不知道，所以我就得带着巧克力棒穿过整个城市。这是暑期最好的工作了。不过我有免费的电影票就是这样好吧，所以我说电影院的票房收益率其实很低电影院的电影产量并不高，因为电影院要花钱租电影放映。所以，你买票的钱很多都用来租电影了。这也是为什么小型电影院只放映数量有限的大片。这些大片的放映成本非常高，所以他们主要通过巧克力棒来赚钱，这一点我已经说过很多次了。所以，去小电影院看电影时，请买巧克力棒。这就是为什么这些数字很小。这就是我想说的。好了，那么我们该如何描述影院的票房和盈利之间的关系呢？我们首先要计算平均值，所以你可以看到 X 栏是 X 的平均值，那么门票销售额 Y 栏就是影院盈利能力的平均值。然后，如果你还记得所有这些表达式，我们将分别计算它们。这是将它们相加的最简单方法。因此，我们创建一个表格，然后分别计算分子和分母。例如，对于 29、49、89 等这些数值，我们用这些数值减去平均值，得出每个数值与平均值的偏差。对于每个 Y 值，我们也用该值减去其平均值的平方。最后取这两个值的乘积。这两个项的乘积。这样，我们就为所有数值创建了一个表格。这并不是很多，每个人只有五个。所以相对来说是可行的。然后，我们就可以计算出我们的贝塔帽子 1，结果是 0.1481，我们的贝塔 0，结果是 14.179。现在，如果我们把这些放到一个奇异的方程中，看起来就像贝塔对不起，y 等于 14.179 加上 0.481 乘以 x 再加上剩余的误差项。因此，你实际上可以把电影院的盈利能力计算为 14 之间的等式。这就是你的截距。因此，你的截距基本上是外部的。由于某种原因，影院目前正在盈利。然后根据票房的高低，会有另一个因素增加利润。更好的情况也可能是负数。在这种情况下，每卖出一张票，电影院就会亏损。好了，现在让我们来看另一个例子。假设我们从美国收集了 209 个工资和公司盈利数据样本。如果我们有了前面的公式，就可以计算出我们的贝塔零和贝塔一，我们就可以在这里找到这个回归模型。销售估算薪金的计算公式为：963 加 18.5 倍，即你所在公司的利润率。该公司的利润率。那么问题来了，一个高管的预测薪酬是多少？B 公司的盈利率为 0%。963.2.没错这基本上就是外部价值了所以总有一个值是 963。现在，我们在 Excel 或 R 或 Python 或其他语言中运行我们的模型，结果发现这个模型的 R 方为 0.0132。谁能告诉我这个值意味着什么？请举手。听两三个方向太难了。我知道就像回到学校一样是啊所以，我们只有1. 32％的变化英寸没错所以只有一个。所以，如果我们想用百分比来表示，我们可以移动逗号，1.3%的模型或工资变化是由这个模型解释的。因此，有很多事情并不属于盈利能力或外部值的一部分。那么，这个数值没有告诉我们什么呢？但它们只是在领带边上被拾起，你知道吗？还有其他变量吗？是的，它基本上没有告诉我们模型中没有的东西。对不对？它基本上是在告诉我们，好吧，如果我们只有工资，我们可以解释其中的 1.3%。如果我们只有芹菜的盈利能力，我们可以解释两者之间 1.3% 的关系。至于还有什么其他因素，我们就无从谈起了。我们也无法说明某些公司的价值是高还是低。这就是我们的想法。如果我们只看整体契合度。例如，我们不知道该模型是否对东海岸的公司非常有效，但对西海岸的公司却完全无效。因此，这只是美国的整体模式。我们也不知道这个模型是否适用于其他国家。因此，我们只有美国的数据来创建这个模型。对于其他国家的数据，它的效果可能更好，也可能更差，甚至完全不适用。这取决于两者之间的关系是否相似。现在好了。假设我们不必手工计算线性回归。假设我们使用某种软件。我很确定这是 R，所以这将是运行线性回归后得到的输出结果。R 在 Python 中看起来是一样的。你明天做教程时就会看到。是的，明天你有一台电脑，明天电脑的输出看起来非常相似。所以它会给我们系数。这里的常数指的是贝塔零，而 x 系数指的是贝塔一。它还会给出这些系数的标准误差，这非常有趣，因为它告诉我们这个系数在数据中的变化或分布情况。我们可以看到，例如这里的 x 系数 beta 值的标准误差非常小，所以变化非常小。在整个数据集中，它始终是一样的。所以，是的，请记住这些都是估计值。我们不知道真实值。问题是我们是否会得到另一个值，在另一组观测中得到不同的值。另一个值是什么？是的，系数的准确性。这一点相当重要。在这里，我们通常会报告一个置信区间。因此，这基本上意味着，如果我们在不同的数据集上重复该实验，我们仍能得到类似系数的可能性有多大？例如，你可以看到这里，我们估计的贝塔零系数是 14，我们有 95% 的置信度认为它将介于 -6 和 35 之间。因此，这基本上给了你一个范围，即使你用不同的数据重复这个实验，你也可以预期这个值会在这个范围内。我们还有一些非常有趣的东西。我们获得了关于假设检验的信息。假设检验将在明天的数据分析原理中深入介绍。是的，在本讲座中也会用到，主要是为了寻找系数的显著性。因此，在本例中，我们感兴趣的是预测因子和响应因子之间的关系是否显著。我们有一个零假设。因此，我们的思维过程是，我们的默认值基本上就是我们的关系。关系不显著。因此，我们更好的假设是零，然后我们的替代假设会更好。在统计上，1 与 0 有明显不同。这就是我们要测试的。希望你还记得一些关于假设检验的知识。这一次使用的是 T 检验，因为这是一个样本，等等等等。因此，我们正在寻找 T 统计量。这是你的 T 栏，我们正在寻找 P 值，如果你们都能记住，希望 T 统计量高，P 值低，因为 T 统计量基本上是在告诉我们，我们的数据是否符合我们的替代假设，这意味着 D 预测因子或贝塔值是显著的，或者基本上不一定很大，但其影响是显著的。因此，你希望 T 统计量非常大，而 P 值则告诉你是否应该拒绝零假设而支持你的备择假设。因此，这里说的是 P 值小于 0.05。这其实是一个经验法则。你可以选择任何你喜欢的 P 值。学生们经常问我，我必须总是用 0.05 吗？不，你可以任意选择。老实说，理论上你可以选择三个 P 值。当然，这并没有什么意义，没有人会真的这么做。但你可以这么做。P 值是你选择的临界值，低于这个临界值，你就会说我认为这是足够的证据。你做出判断：这是足够的证据来拒绝我的零假设。这只是一个指标，基本上是给自己一个必须遵循的临界值，可以是 0.05，可以是 0.01，也可以是 0.1。这三个值是最常见的值。因此，在这种情况下，我们可以说，例如，我们可以看到这里的 P 值是 0.08。因此，如果我们选择的临界值是 0.1，我们就可以说该系数在 10%的显著性水平上是显著的。你也可以看到它不小于 0.05，我不希望你说 "几乎显著 "或 "接近显著 "之类的话。我会把论文烧掉的。不，我不会那么做的这太过分了但如果我发现根本不存在什么接近重要或几乎重要，我会非常生气。就是没有我们刚说了这是你的阈值你来决定。你说，这是我的水平，要么在这个水平之下，要么不在这个水平之下。要么是，要么不是。没有中间值。我见过这种情况。我有时会在论文中看到这种情况。因此，我认为这实际上是研究中的一个大问题。我希望明天你们能听到我的同事唠叨，说人们其实一直在追求显著性，但实际上并没有这种东西。是有意义还是无意义，二元对立。好吧。我们已经讨论过这种简单的线性回归。现在让我们简单看看多元线性回归，概念非常相似。我们只是将其扩展到包括一个以上的预测因子。因此，我们使用简单线性回归，现在我们包含的不是一个预测因子 X，而是 K 个，K 可以是你选择的任何数字。因此，根据你的数据集大小，我们仍然有相同的成分，我们仍然有截距，我们仍然有误差项 Epsilon，在最末端。所以，我们背后的逻辑还是一样的。我们仍然在寻找你的系数估计值，也就是你现在所有的贝塔值。因此，我们会尝试对任何实例 X 及其值 y hat 进行预测。想法是一样的。我们只是把它扩展到了更多。现在，你也可以用矩阵符号来写，如果你用数据集和数据帧的术语来思考，这可能会非常有用。举例来说，如果你的 Excel 电子表格中包含所有数据，那么所有数据就是所有 X 值。例如，你有 N 个观测值，200 人回答了你的调查，你问了他们 10 个问题。那么 K 等于 10，等于你说的什么？100?200?我不知道。你的申请数量。我只是随便告诉你一些数字 然后就忘了我应该写下来真糟糕 That's terrible.好了 现在我们有了贝塔向量 Okay, we have our vector of betas now, 所以每个预测变量都有一个更好的贝塔 so we have one better for each of our predictors, for each of our variables.K 每个观测值都有一个误差项。这样就有了误差向量、系数向量和矩阵，基本上就是数据框架了。我们也采用同样的方法来找到误差项。因此，我们仍然使用最小二乘法，只是稍微改变一下计算方法，将所有不同的 X 和所有不同的投注者都包括在内。因此，我们实际上要计算的是 K 个下注者。我们仍然可以计算一阶条件，以最小化我们的平方和，并求解贝特值。现在我们可以看到，我们的贝塔帽，记住这是一个向量，实际上可以这样计算，现在你有你的转换矩阵 X 乘以你的矩阵 X。取其逆。等等。等等。这样就得到了贝塔帽。我的同事认为这是个很好的记忆方法我觉得这真的很可爱。所以请记住，最好的帽子是蓝色的，因为这是最好的线性无偏估计器，基本上它是你对线性模型的最佳拟合，它是无偏的，是你的估计器。我不知道你是否觉得这对你有帮助。我觉得这是个非常好的花絮。所以，请记住，如果你使用最小二乘法，你是在试图为该数据集找到一组最佳的击球手来预测你的 Y 值集。好的现在我们还有一点时间。我希望你们现在能清醒一下，因为今天已经很漫长了，所以我希望你们能根据我们今天在讲座中讨论的内容思考以下问题。但你们也可以利用自己的经验，因为我知道你们中的很多人都有分析的背景，所以你们过去一定用过线性回归。想一想线性回归模型的几个优点，想一想其中的一些缺点。你可以先自己想，也可以和你的邻居或三四个人的小团体讨论，我并不介意。然后在 5 到 10 分钟，比方说 7 分钟的时间内，我们将与全班分享，收集其中的一些优缺点。我之所以要告诉大家，或者说为什么要让大家这样做，是因为在这个系列讲座的一开始，我就强调了 对我来说，重要的是要认真思考什么样的模式是合适的，在这种情况下，每种模式的优缺点是什么。因此，在介绍完每种模式后，我们将进行类似的讨论。最后，我希望你们看看自己的笔记。我建议你们写下笔记。为什么要讨论这个问题，为什么你的同事会这么说，因为这有助于培养你的批判性思维能力，以便能够批判性地讨论这些模式并对它们进行比较。好的，花几分钟时间，然后我们将与全班分享。和全班分享总是让我听起来像个中学老师。很抱歉让我们重新开始录制。谁想开始？你可以开始优点缺点，或者只是一般性的意见，你不完全确定是否符合这些类别中的某一类。我们会大致走一遍房间，这样我就能听到每个人的意见。我们从这边开始，前面那个。你们讨论过吗？你们知道的。我们讨论过的一个优点是，任何人都可以对它进行解释。因此，如果我想编写代码，它非常容易理解；如果我创建了一个模型并将其展示给其他利益相关者，它也非常容易理解。他们很容易就能理解我们实际上做了什么，但我们认为，基本上我们认为它简单易懂，但有时将它应用到复杂问题上非常简单，也许我们的解决方案并不符合要求，因此我们认为它是一个缺点。它有时会为非线性问题提供非常线性的解决方案。因此，从根本上说，它是简单的，但也可能是简单化的。而且，它不能为我们提供所需的问题的正确解决方案。基本上是这样是啊。所以，你应该还记得，在第一讲中，我们在讨论整个系列讲座中会涉及哪种模型时，我们基本上是在不断增加复杂性。所以，我们将从线性回归开始，简单的线性回归研究，非常简单，正如我所说，易于理解，易于解释，易于向别人解释，因为它是一种简单的线性关系。我的意思是，我们正在讨论这个问题，例如，销售额和利润率之间的这个例子，或者一些有 1 或 2 个输入或几个输入，而你有一个输出的例子。你可以解释你的因素、你的参数、你的赌注。你可以非常简单地解释，如果增加 10 个单位，就会产生确切的数字影响。但你是对的，这可能很简单。这是一个线性模型。我们稍后会看到更复杂的非线性模型。例如，神经网络就非常复杂。它们基本上处于复杂度的另一端。因此，对于高度复杂的关系来说，它们是很好的选择，但在解释和说明实际情况方面，它们却很糟糕。所以说，两者都有。你必须平衡这两者。很好在那组的后面很好在后面，在边上你们讨论的是优势还是劣势 What have you discussed can be advantage, disadvantage.一般的评论，因为我怎么能说可以通过。把一个单独的点放在一条线上，它可以让你看到非常非常简单或非常刁钻的观点。情节，你可以看到的趋势。所以这是一种可视化的优势，你也可以将结果可视化。是的，我认为这是一个很好的优势。这也是可解释性的一部分，因为你可以比较容易地显示出其中的关系。我们刚刚还讨论了片断线性回归的问题。因此，即使关系总体上是线性的，但可能会分成几个步骤，那么你仍然可以使用线性方法来处理。举例来说，如果你把数据空间划分为不同的部分，然后你可以把线拟合到数据的不同部分，并对其进行解释，我想我被你的一位同事问到过这个问题。举例来说，如果你在 X 的特定区域有一种线性关系，而在 X 的不同区域又有不同的线性关系，例如，年龄与收入之间的关系，当你的客户样本年龄增长时，你的收入会普遍增加，然后你突然有了一个更高的阶梯。因此，在这个年龄点上发生了一些事情，我们的收入突然增加，然后又是这种一般的线性关系。举例来说，你可以用 "哦，这个年龄段的人正在升职 "或类似的话来解释。所以，你也可以尝试在片断线性回归中为这些步骤找到解释。是的，非常好。你们在一起工作时，从后面或从侧面看，我们还有其他优势或劣势吗？很大？是的这个模型真的很强大，所以它很简单。几乎所有的金融应用都会用到它，比如模式建模。我们喜欢会计位像预测风险 或像排序的股权在企业社会责任和东西。它真的很强大。我的意思是，它几乎可以应用于各个领域。非常简单。鲁棒性是这样吗？是的这是一个大问题。我觉得这又回到了我们之前在讲座中讨论的方差偏差权衡问题。因此，一个简单的模型可以非常稳健，非常容易应用于很多不同的场合，正如你所说的那样。正如怀特所指出的，我们之所以首先提到这一点，是有原因的。这是很多建模工作的核心。因此，在很多情况下，你会把它作为基线模型，或者把它作为第一种基本方法进行比较。在很多情况下，这就是你所需要的一切。有时你并不需要复杂的模型。有时，一个简单的模型就能完全满足解释关系的需要，而且可以在很多不同的情况下使用。好了，优点。缺点。在中间的某处来回穿梭。对我们说过它只适用于数值变量变量不能用别的分类变量作为结果处境不利？是的。作为输出或你不能。这只是数字。哦，是的。是的，输出，是的。所以你只能预测数字输出。数字输出我的意思是，你可以使用分类变量作为预测变量 如果你，例如，使用一个假编码或我们使用的Simula。可以吗？是的 是的所以输出必须是数值型的。下周我们讨论逻辑回归时就会发现，我们可以将分类输出作为变量。所以我们会发现有一些方法可以解决这个问题。但对于我们今天讨论的简单线性回归，它的结果必须是数字的。没错没错提到你必须假定自变量或预测因子不会相互影响，它们是独立的。是啊所以你要假设预测因子是独立的，它们之间不存在共线性。举例来说，如果预测因子之间存在完美的共线性，模型就会完全崩溃。很多学生在尝试线性回归时都遇到过这种情况，比如，他们创建了所有的虚拟变量，却没有删除其中一个。因此，这些虚拟变量是完全相关的。我就告诉他们，不，我不会这么做。然后他们来找我，问我为什么不行？有什么问题吗？问题在于，如果变量完全相关，就会完全崩溃。而如果你有强相关变量，那么你的输出就不可靠了。因此，如果预测变量之间存在相关性，你就不能真正相信线性回归的结果。因此，我建议在使用线性回归之前，首先要做的一个重要步骤就是创建相关矩阵。看看你所有的独立变量，是的，你所有的独立变量之间的相关性有多大。我想你会在实验室里做的，或者你可以在实验室里做。所以，提示提示。我明天的实验方案包括创建相关矩阵，你们也可以看看。这真的很有用，也是很好的第一步分析，在这一步中，你不仅要研究自变量与因变量之间的关系，还要研究数据集内部的关系，然后有办法克服高度相关的因素。例如，你可以使用降维技术，创建因子或主成分。