题目：

有一个250人的语音库，共16万个音频，每一个音频都有与之对应的转写文本，音频和文本数据按人存放。 语音库中有男女、发音人地域、每个音频时长等统计信息，也有借助自动语音识别系统得到的每句音频的近似字错误率信息。

要求通过统计学的方法统计该数据库的转写正确率

（即文本和语音是否一一对应）。

**请给出自己的解决方案及理论依据。**

一：问题分析

数据库的转写正确率是与音频对应的发音人的性别、发音人地域、音频时长有关的。男性声音粗而低，女性声音则细而高，那么在语音识别时对于男性和女性识别正确率可能是不相同的。其次，发音人所在地域不同，方言也就不同，甚至同时讲一样的普通话发出来的音调也是千差万别的，因此发音人地域也在一定程度上影响转写正确率。最后，音频时长可能对转写正确率也有一定的影响，音频时间越长，那么语音识别系统识别犯错的可能性就越大。

根据题目得知：观测的样本数据有16万个（样本量很大），观测的变量有性别、发音人地域、音频时长和错误率信息，要求是统计数据库的转写正确率。转写正确率在统计学意义上相当于,其中表示正确率，也即1- 错误率信息，则表示影响正确率的协变量，表示相应的观测值。显然为的函数，这里记为。我们的目的就是根据已观测到的样本数据来确立的形式，或者给定的值我们可以得出对应的转写正确率。

二：解决方案及理论依据

**方案A及理论依据**：

假设为最简单的线性形式，即。其中是要估计的未知参数。参数估计方法我们可以采用最小二乘或者稳健的估计方法（当观测值存在异常值时）。然后我们需要评价所建立模型的好坏，具体的做法可以利用一些评价指标去衡量，比如残差平方和、AIC、残差平方图等。我比较常用的方法是交叉验证，即将观测数据分成两部分训练集和测试集，然后计算CV值，根据CV的大小来判断模型是否合理。

如果模型不合理或者不满足线性假设，可以进一步利用非参数平滑的方法求解。非参数平滑法包括很多方法，比较常用的是局部多项式拟合，原理可以简单地理解为在局部进行线性回归。由于协变量是多元的，在利用非参数平滑法求解时，我们可以将设为的形式，即单指标模型。对于单指标模型我们需要估计的是和函数。单指标模型的算法我们可以采用迭代的算法计算。同样可以用CV评价模型的优劣。

理论依据：

1、最小二乘法估计即最小化。

2、单指标模型对于和函数的估计如下：

为了参数的可识别性，对施加约束。可以将上述线性回归的结果作为的初始值，对于可以利用任何非参数平滑器比如局部线性回归平滑估计。

**方案B及理论依据**

可以应用回归树模型来统计数据库的转写正确率。具体做法如下：和分别为输入变量，将上述语音库记为训练数据考虑如何生成回归树。

一个回归树对应着输入空间的一个划分以及在划分的单元上输出的值。假设已将输入空间划分为个单元，并且在每一个单元上有一个固定的输出值，于是回归树模型可表示为



当输入空间的划分确定时，可以用平方误差来表示回归树对于训练数据的预测误差，用平方误差最小的准则求解每个单元上的最优输出均值，即



问题是怎样对输入空间进行划分。可以采用启发式的方法，选择第个变量和它的取值，作为切分变量和切分点，并定义两个区域：

和

然后寻找最优切分变量和最优切分点。将输入空间划分为两个区域。接着，对每个区域重复上述划分过程，直到满足停止条件为止。这样就生成一棵树。根据所得到的这棵树来统计该数据库的转写正确率。