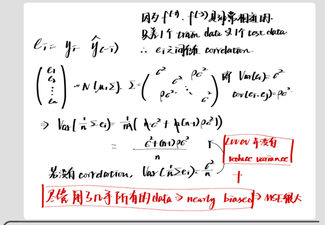
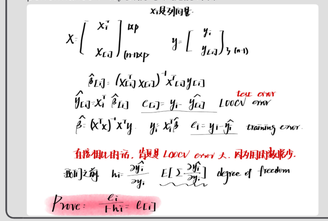
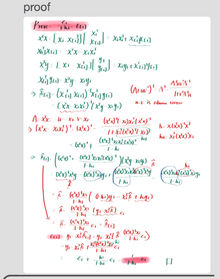
- Resampling
  - 当我们进行sample的时候，我们确实是可以得到population的某些信息，从population中抽取多次样本，这些样本的均值，这些均值放在一起就是sampling distribution，把每个sample当成每次训练数据进行linear model拟合，从而得到一个$\hat{\beta}$，很多个$\hat{\beta}$可以求出β 的均值和方差
    - 但这个在实际操作中非常困难，很多时候我们都只有以一个样本，也就是一个训练集，因此我们得不到β 的很多个估计值，那怎么办呢？
    - bootstrap
      - bootstrap的意思是：自助，也就是不需要依靠其他的东西，而是仅仅是这一组数据就可以解决问题
      - 怎么解决呢？就是我们把我们唯一有的这个数据（所以为什么把这个数据当成population，因为我们只有这一组数据）当成总体，然后模拟从总体中抽取样本的过程
      - 这也就能解释，为什么我们要抽取n个样本（因为我们要模拟抽取n个数据的样本的过程）
      - 之后，为什么要有放回的抽样，我们设想，如果不放回，我们抽取了第一个样本，和原样本一模一样，在抽取？就仅仅是复制样本一，这是没有意义的，所以要放回，同时放回保证了每个数据被抽取到的概率都是相等的
      - 这样我们可以获得M个beta，这些beta模拟的是我们原样本的均值，但他的方差可以拟合我们beta，也就是真实值的方差，这种resample的beta成为bootstrap distribution
        - bootstrap distribution的variance可以approximate sampling distribution
        - relationship between bootstrap dustribution and sampling distribution
          - structure part is same, but not the mean part
        - was used for statistic inference
        - data pertobation?
          - add some changes to data-->a model
          - add another changes --> another model
          - 可以帮助得到一些robust，但是不能理论上的证明
  - 我们有一个数据，需要将其人工分为train data & test data
    - 那就会产生两个问题
      - 怎么分，是 五五分，还是三七分？
      - 怎么选择数据，这是有randomness的
  - Validation set approach:
    - 比如我就分为50%50%,每一个train data随着自由度变化，可以画出一条MSE curve，那么如果我进行resample10次，将MSE取平均，就能够一定程度的降低randomness
    - 那proportion怎么解决呢？如果更少的数据，那么我们的模型肯定是更不准确的
  - LOOCV
    - 每次选择一个数据点作为test error，其余的作为train data,因此模型的拟合只有n次，最终的MSE就是n次拟合的MSE的均值，这样就完全消掉了randomness
    - disadvantages
      - confidence interval 偏小<--由于model is highly positively correlated
      - point estimate is good
    - variance偏大
      - 
    - Modified<--我们可以通过一次拟合就可以得到LOOCV的performance<--h_i 我们可以提前计算
      - 
    - proof
      - 
    - 同样的，Ridge Regression 也是这样的，不需要一定做n次拟合了
  - K-fold cross validation
    - 将整个数据分成K层，一般K=5/10,，每次将一个K分出来，作为test data，其余的作为train data
    - 他为什么有的时候比LOOCV效果好？因为Variance小
  - How to combine variable selection and cross validation
    - experiment: 我们设计50个variable，都独立服从于t分布，而Y服从于伯努利分布，也就是说，X和Y是independent，当我们训练一个分类器，他的正确率应该是在50%左右才是正确的
      - 正确的是我们应该先分train data & test data,对train data 进行variable selection，也就是说，我每次variable selection 的数据都不一样，这个必须不能涉及test data Y，我不在乎你到底对train data做了什么，什么都可以，也就是说cross validation必须包含variable selection和训练一起
        - if 我们先用每个单独的X和Y做回归，寻找p最小的20个variable，然后利用这20个X变量和Y进行训练分类器，进行cross validation，我们发现这个分类器的正确率在14%，远远低于50%，为什么？
        - 因为我们犯了data snopping 的错误，我们已经看了所有的X和Y数据了，但又假装我们没看过，然后去进行训练和cross validation，也就是说，我们在期末考试前已经看了考试题和答案了，那我根据这个学习，再去做期末考试，我的成绩肯定会远远高于没看过期末考试答案的，而train data相当于去年的期末考试题
      - 那么，如果是PCA呢，两种做法会有不同吗
        - 不会，因为我PCA只看了X，并没有关注Y，哪怕看了test data X，但你只是看了期末考试题，不会就是不会