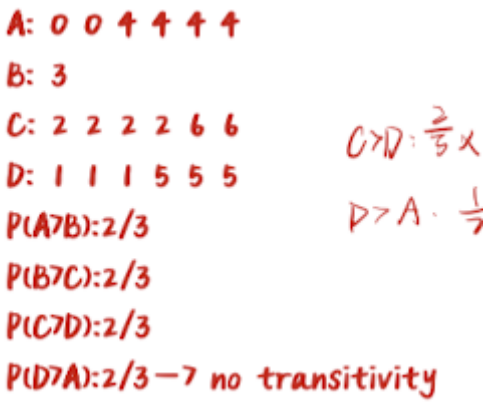
# Classification

- uncertainty:
  - deterministic: when X>Y, Y>Z --> X>Z
  - but in probabilistic: we consider 4 dices:



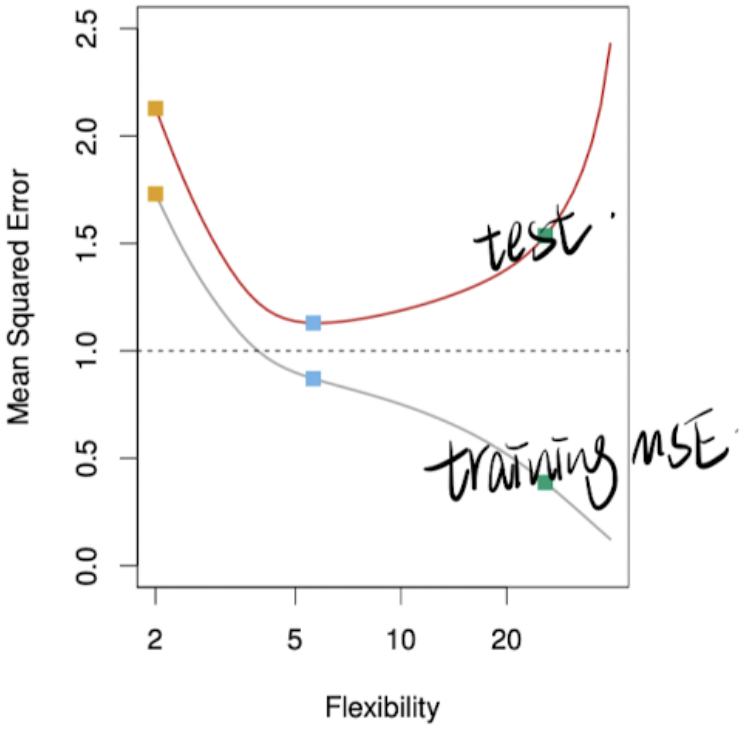

- Two methods:
  - Bayes inference: only know the past data and distribution ,and we can get new information, then we can updata the distribution, e.g. weather
  - Frequency: on the same condition and repeat many many times
    - pros: easy to understand
    - cons: only suitable for some special cases
- Statistical machine learning
  - $Y = f(X) + \epsilon$
  - reansons:
    - prediction
      - error
        - reducible:
        - irreducible: Y is also a function of $\epsilon$ , this can't be predicted by X
    - inference
  - estimate f
    - Parametric
      - procedure
        - assum the functional form
        - estimate parameters
      - cons: assumption may bot correct
      - how to solve this? choose flexible models<- but more complex model can lead to overfitting

- Non-parametric： 不是没有参数，而是没有预先设定的参数形式
  - cons: require a lot of data(far more than parameter model)

- Assessing model accuracy
  - measure the quality of fit
    - MSE: most test MSE(注意没有irreducible error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

      - plot




        - when training MSE decreases, testing MSE increases --> overfitting<--think some random things as the reason.
    - smoothing spline

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

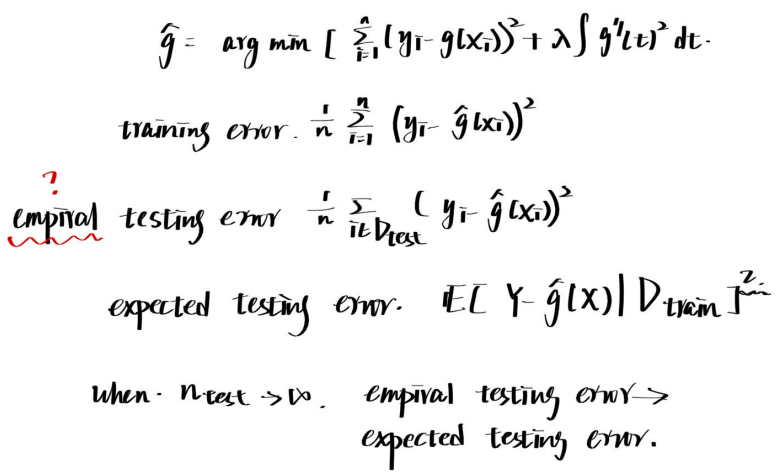
      - $\lambda \to \infty$, 不能有太多的二阶导的影响，所以是linear function
      - $\lambda \to 0$, more flexible function
      - 根据曲线balance between underfitting and overfitting

    - the smoothing spline seems to get min testing MSE, but why we still need some other methods? In principle, it's perfect, but in practice, it's not perfect especially in the high dimensional case.

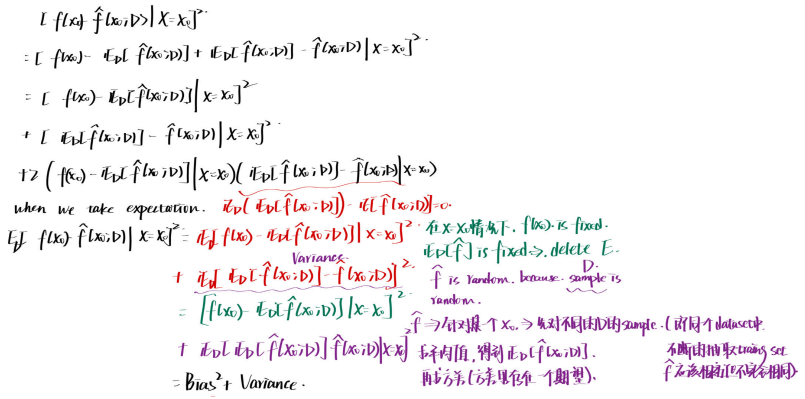
- The Bias-Variance Trade-Off
    - formula

$$\min_{\hat{f}} \left[ f(x_0) - \hat{f}(x_0; \mathcal{D}) \middle| X = x_0 \right]^2 ?$$

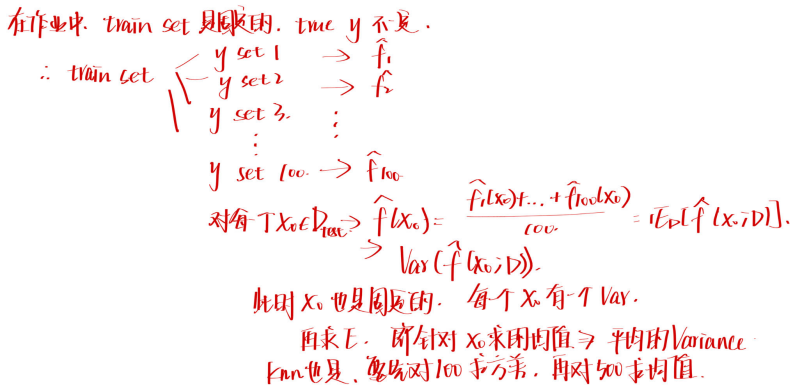
        - $\hat{f}$ 是根据training data 训练出来的，$x_0$ 不是training data里的数据（因为我们想要的是min test MSE），且我们不需要一个只适用于某一个特定dataset 的模型，而应该是最好所有的dataset（training set）， 在这个想法基础上，我们需要加上期望
        - MSE



            - 事实上，有没有条件是有一定区别的，但非常非常小，可以忽略，因此我们不考虑是否存在condition
            - 化简



                - 思考



- 这里的bias 指的是，在固定$x_0$的情况下，我用的模型和真实值差多少
- Variance指的是，在固定$x_0$的情况下，当我用同一个大的dataset，但是不同的training set的情况下，$\hat{f}$ 也就会变，他的变化浮动大不大，也就是对不同的training set 的敏感程度如何，
- 例如： a constant model, f=c, 不同得training set得到的model相同，variance=0，但bias会非常大； a very flexible model, 对training set的敏感程度很大，variance很大，bias会比较小<--因为比较准确

- plot


- Why machine learning method is flexible but small error?
  - 1. signal +noise, signal is very large than the noise--> unlikely to overfitting
  - K line?
- 不是说如果true is non-linear 就不能用linear model to estimate, 如果bias最小，那么也是可以用的

- linear model
  - simple linear regression
    - concept

population level: $Y = \beta_0 + \beta_1 X + \varepsilon$.

↓ sampling ← frequent perspective.

sample level: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

$\langle \hat{\beta_0}, \hat{\beta_1} \rangle = \operatorname{argmin} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$.

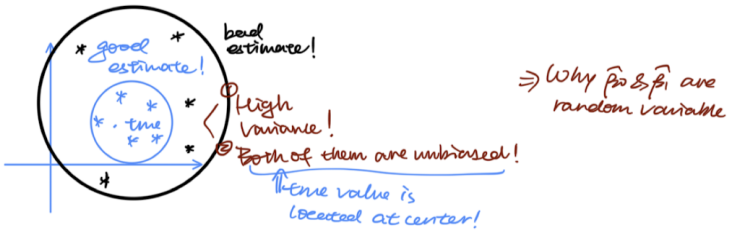
$(\hat{\beta_0}, \hat{\beta_1})$ is the realisation of $(\beta_0, \beta_1)$

$\langle \hat{\beta_0}, \hat{\beta_1} \rangle$ is random variables depending on sample.

- LSE: least quaies criterion
  - $RSS = \sum e_i^2$
  - coefficient estimate
    - The least squares coefficient estimates are given as

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

      where $\bar{y} \equiv \frac{1}{n} \sum_i^n y_i$ and $\bar{x} \equiv \sum_i^n x_i$.
  - Assess the accuracy

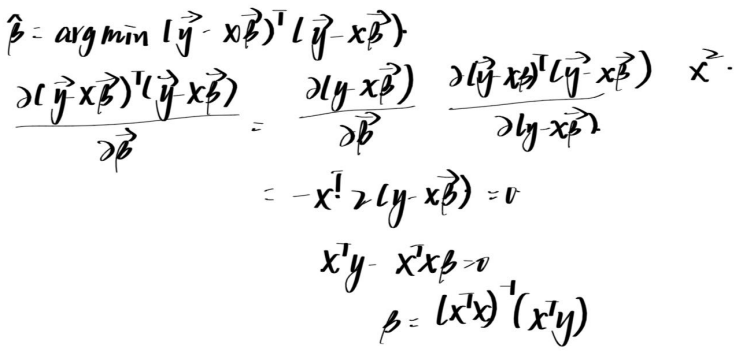



    - SE: 标准误，一般等于$\frac{\sigma^2}{n}$,在这里，为什么减2，因为是2个coefficient，参见LCM

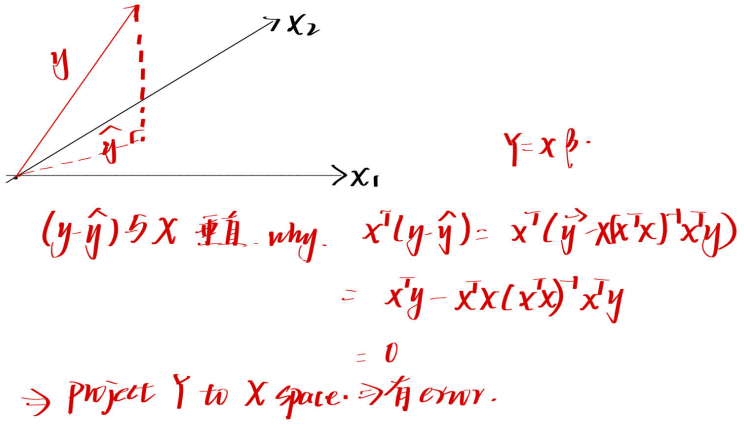
$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_i^n (x_i - \bar{x})^2} \right], SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_i^n (x_i - \bar{x})^2}, \quad (7)$$

      where $\sigma^2$ can be estimated by $\hat{\sigma}^2 = RSS/(n-2)$.
    - 注：我们可以说$P(\hat{\beta} - 2se \le \beta \le \hat{\beta} + 2se) = 0.95$,但如果我们一旦知道了$\beta = 1.1, se = 0.2 \to P(0.7 \le \beta \le 1.5) = 0.95$ 这是不对的，因为此时，P内的数字都是固定的，不是random variable, 意思是说，我们有95%的把握认为真实参数是在这个区间内，而不是真实参数在这个区间内的概率是95%
    - $RSE = \sqrt{\frac{1}{n-2} RSS}$ residual standard error
    - $R^2 = \frac{TSS - RSS}{TSS}, TSS = \sum (y_i - \bar{y})^2$
      - for training data, non negatve
      - --> 1 better
      - onlu useful for simple linear regression
      - just focus on training error

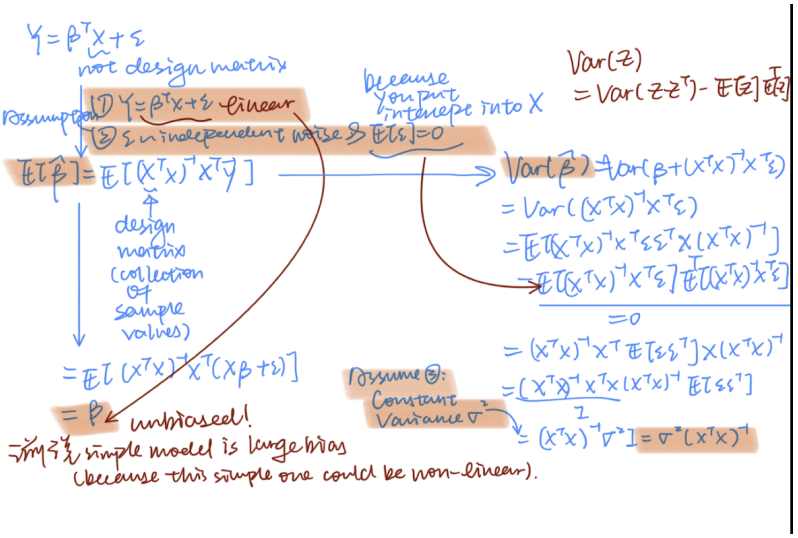
        - for testing data it may be negative <-- overftting
- multiple linear regression
    - concept: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$
    - estimate
        - method 1



        - method 2
        - understanding using projection



        - 在assumption下，unbiased and $Var(\hat{\beta})$ 固定

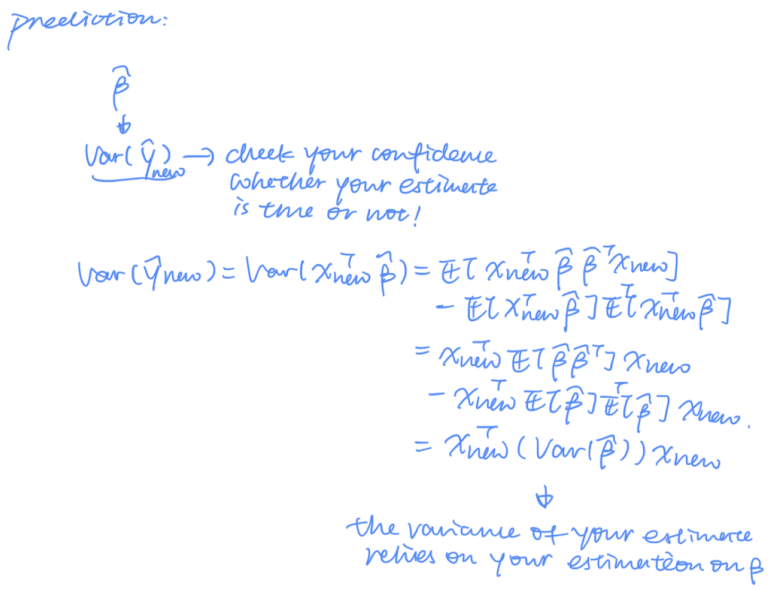


            - 之前我们说linear 的bias 大，是因为不一定原模型就是linear model

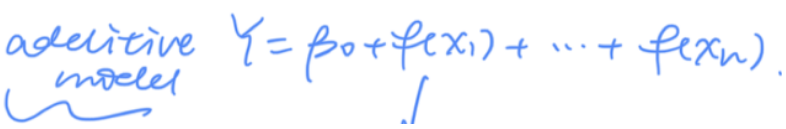
- Is there a relationship between X and Y?
  - F test: $H_0$ : all the $\beta = 0$

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)},$$

- Deciding on importance Variables
  - T-test and model selection is different, T-test是想控制type1 error， 也就是在原假设为真但我拒绝了，model selection 是想找出预测最准确的model(focus on testing error)，两个给出的结果不一致，但在假设正确且数据足够大的情况下，结果会趋于一致
  - 我记得LCM说的是：AIC，BIC都存在E，但我们只用了一组特定的数据来计算Cp，所以结果可能不一样
- Model Fit
  - 可能会发现，变量个数越多，$R^2$ 越大，这是因为我们的$R^2$ 是用training data 来计算的，变量越多，哪怕和谐变量可能并不显著，也会导致我们的training error 越小，$R^2$ 越大
- Prediction
  - model bias(很可能不是linear model)
  - reducible error
  - irreducible error
  - confidence interval



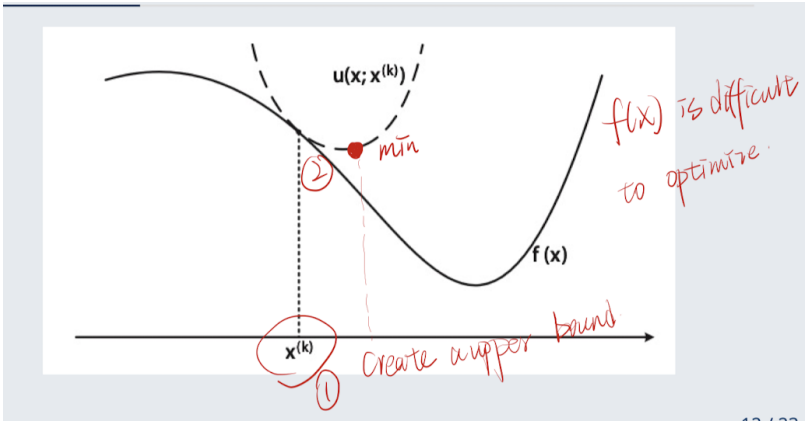

- Extensions
  - linear model 最重要的两个假设: 1. additive 2. linear
  - 什么是additive: $X_j$ 对Y的影响相互独立




- 如果我们remove这个assumption，$\beta_3 ! = 0$ 说明，$X_1$ 对Y的影响还取决于$X_2$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- Potential Problems
  - non- linearity：residual plot，不管维度多少，这个图都非常有用且简单： if residual is not zero--> linear model is not approciate and try to add some nonlinear parts
  - correlation of error terms: $\epsilon$ 不独立，那么$Var(\hat{\beta}) ! = \sigma^2 (X^T X)^{-1}$ ,这个表达式会低估Var， 所以confidenc interval 会偏窄，也就是说，真正的这个区间的把握应该小于95% --> GEE 可能可以解决这个问题
  - non-constant variance of error terms: 我们之前假设$Var(\sigma^2)$ 是不变的，但如果他变化呢？可以变形Y--> log(Y)/$\sqrt{Y}$ （concave function） Y会变小，异方差的影响也就会更小一点
    - 有时一些X的方差我们是已知的，这时就需要修正我们的方差，
  - outliers: 1. largely deviate from other data 2. 数量很少
    - how to solve:
      - method 1
        - $\omega_1 = 1$ --> find outliers
        - change $\omega_i = \frac{1}{|e_i|}$
      - method 2
        - 首先我们用L1代替L2<-- 因为L2 对数据的变化太敏感
        - 但L1不好的地方在于它是non-differentiable 的
        - 所以我们找一个L1 的upper bound 且需要differentiable




        - $x_k$ 已知，找原曲线上该点的upper bound function，找这个function的最小点，--> $x_{k+1}$
      - but how to construct u function?

- formula

$$u\left(t, t^k\right)=\frac{1}{2\left|t^k\right|}\left(t^2+\left(t^k\right)^2\right) .$$

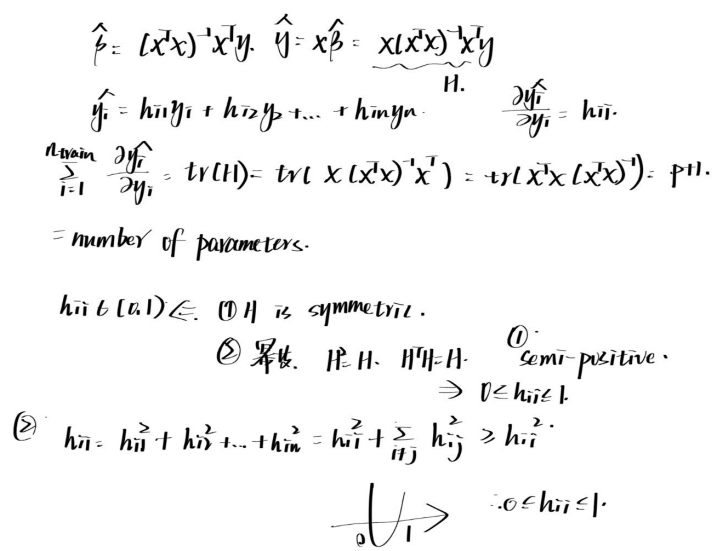
$$u\left(\beta, \beta^k\right)=\sum_{i=1}^N \frac{1}{2\left|[X \beta^k-y]_i\right|}\left([X \beta-y]_i^2+\left([X \beta^k-y]_i\right)^2\right) .$$

- argorithm

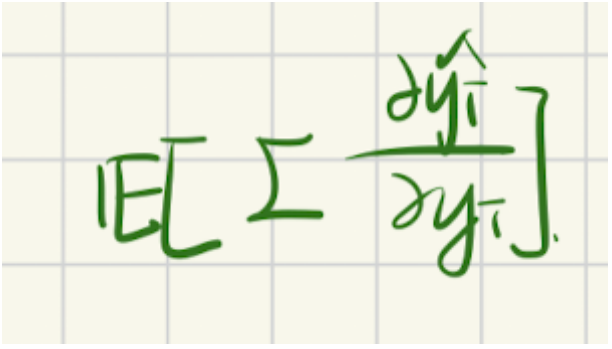
  - Set $k=0$ and initialize with a feasible point $\beta^0$
  - **repeat**
    - $w_i^k \leftarrow \sqrt{\frac{1}{2\left|[X \beta^k-y]_i\right|}}$
    - Update $\beta^{k+1} \leftarrow \underset{\beta}{\operatorname{argmin}}\left\|(X \beta-y) \odot w^k\right\|_2^2$
    - $k \leftarrow k+1$
  - **until** convergence
  - **return** $\beta^k$

- 注意：不能用residual来判断是否有outlier，因为我们的residual plot和我们用outlier预测出来的model是有关系的，有可能他针对元数据就是outlier，但拟合的时候我们不认为它是outlier

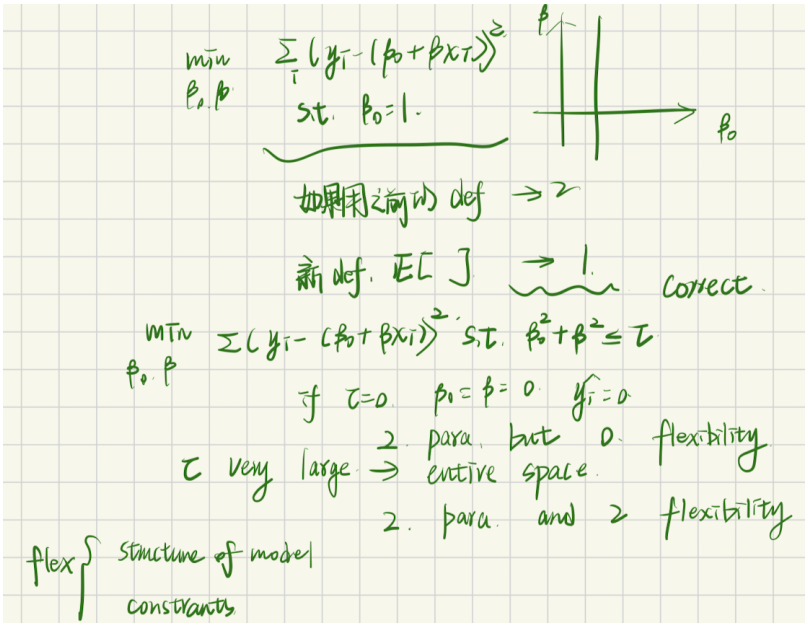
- high leverage points $\frac{\partial \hat{y}_i}{\partial y_i}$ 非常大，也就是这个点对$y_i$的变化非常敏感
  - model flexibility =sensitivity, 即我对于这个点，仅仅换一个training set的y值，我拟合的这个model变化大不大?
  - 猜想--> $\sum \frac{\partial \hat{y}_i}{\partial y_i}$ ,how to prove? (if it can cover the special case, then we think it is correct. eg. linear model's flexibility is the number of parameters)



- flexibility is upper bounded by number of data--> so we need many many data
- but $\hat{y}_i$ is random variable, --> 以上公式random variable, 所以为了获得一个更准确的公式， we need expectation



- example



- Collinearity: eg $X_1 = \alpha X_2$
  - 两两之间： pairs
  - 有的不能直观看出<--multi collinearity<-- VIF : 越大说明越有collinearity

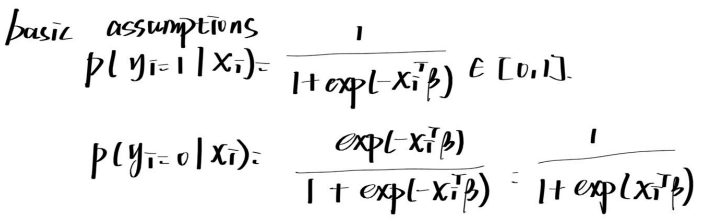
$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the $R^2$ from a regression of $X_j$ onto all the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

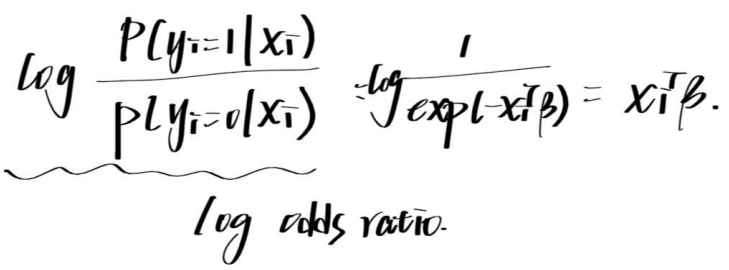
- 但可能没那么好，因为一个一个做我们需要做p次，$\lambda > 0$，这应该是个修正，本来如果共线性，$(X^T X)^{-1}$ 是不存在的，在这加了一个$\lambda$ 让它存在

- VS K-Nearest Neighbors (KNN)
  - non- parametric
- 不能简单认为原数据是non-linear 的，就该用non-linear 拟合

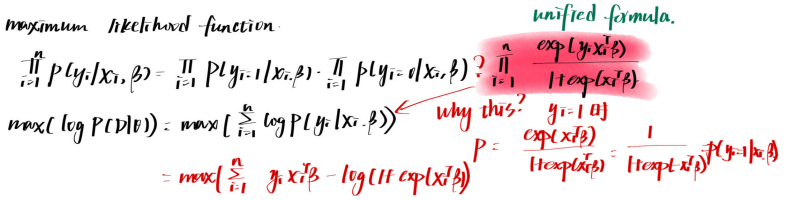
- Classification
  - supervised learning
  - why not regression
    - qualitative response 不能用量化数字来表示
    - 类别有时是有顺序的，例如severe/moderate/mild， 但如果我们用2，1，0来表示，数字说明他们之间的差别是相同的，但实际上不是这样的，我们不知道程度之间的差别到底相不相等，这是无法量化的
    - 我们没办法确保$\hat{y} \in [0, 1]$
  - LDA(linear discriminant analysis)
    - sigmoid function/softmax active function



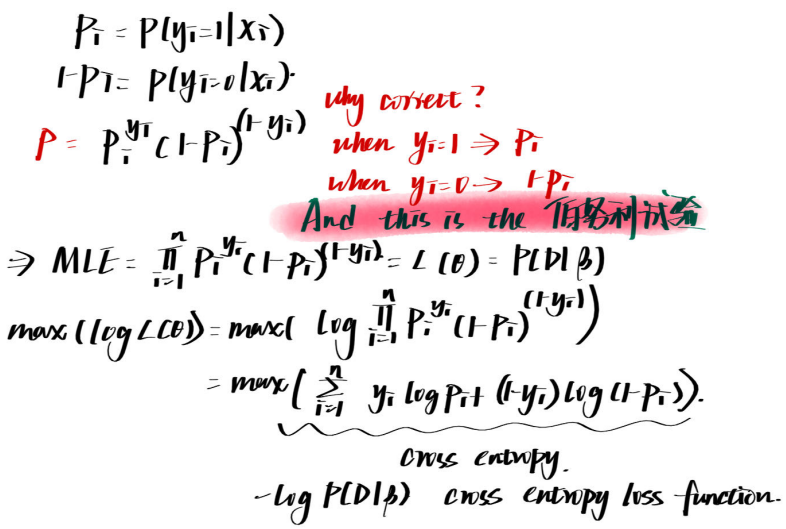
      - assumptions: sigmoid function +$y_i|x_i$ is independent
    - why this function? can we choose another function?
      - Yes, for example $\int_0^x N(z)dz$， z foloows normal distribution-->probit model
      - The performance are nearly the same
    - logistic function: 1. nearly unbaised 2. easy interpretation
    - in fact, it is the generalization of linear model



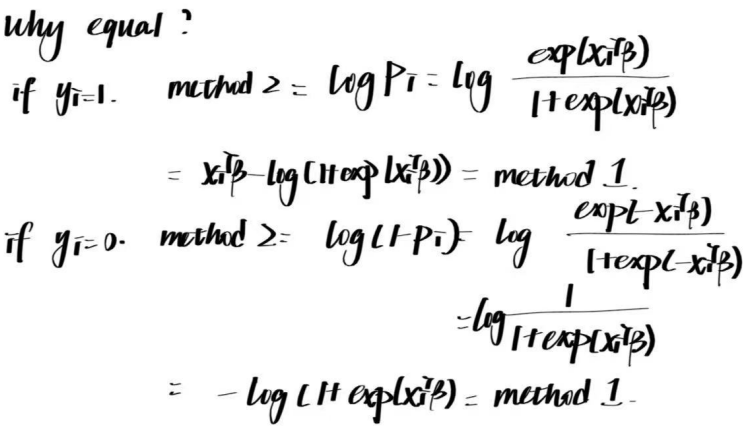

- we have two perspectives to understand
  - method 1



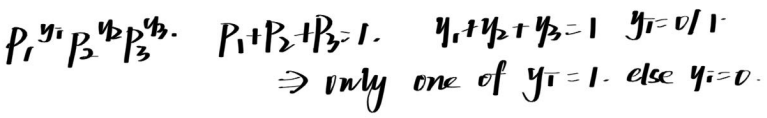

  - method 2



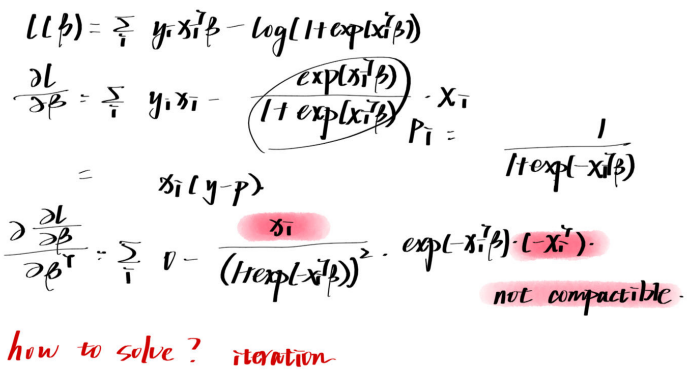

  - why equal?



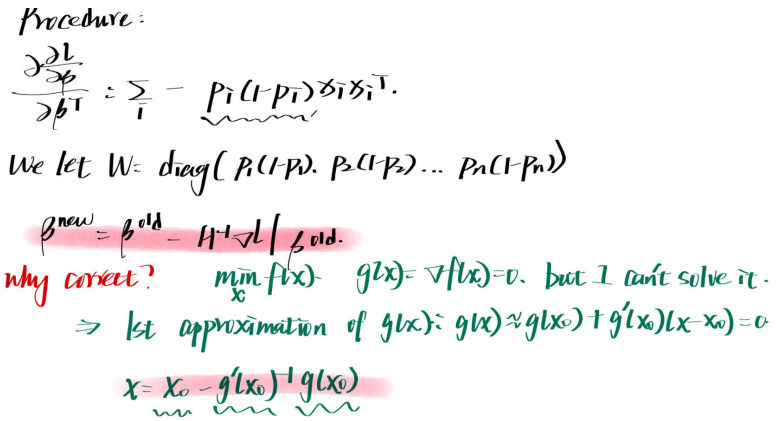

  - why we need the method 2? Because it is easier to generalize



- how to solve the MLE?
  - 一般来说我们求导计算

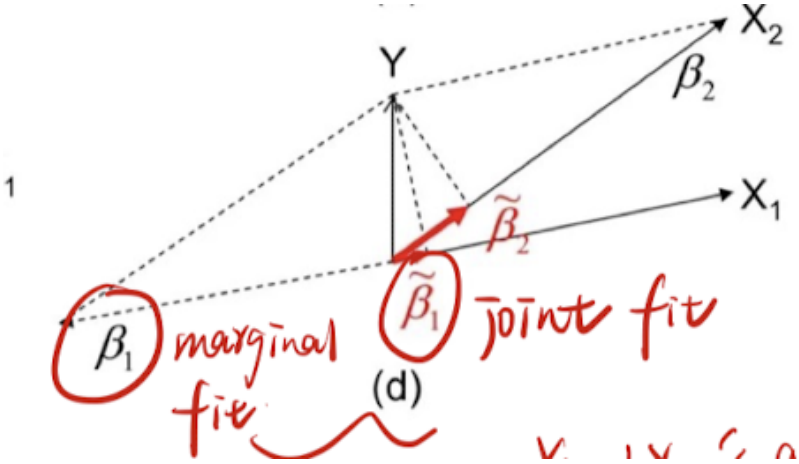


  - 但这里我们不能解决二阶导，how to solve?



  - 我们得到的结果会依赖于$\beta_0$吗
    - 不会，因为这是一个convex problem<-- 二阶导大于等于0

- example and paradox
  - 我们得到的系数都是在有其他系数的影响下得到的系数，也就是影响程度,即joint fit
  - 如果仅有一个X因子，才是不condition on other factors--> marginal fit
  - depending on the angle of these two factors
    - marginal fit 是平行四边形分解，而joint fit 是有Y向两个X作垂线




- 两个系数都正确，取决于我们如何解释这个系数，joint fit即是conditional 的

- Multi- class
  - 我们首先分析一下两个名词:
    - generative model： 什么是生成式模型，就是我在给定这个模型的结构和参数之后，可以直接生成数据，最简单的就是：如果我知道了一个函数f的分布，那么我可以不需要任何别的参数就可以生成符合这个分布函数的数据-->从这个角度来看，logistic Reg不是生成式模型，而是判别式模型, discriminative model(给定了模型的结构和参数，还必须知道X才能生成符合条件的Y，如果没有X，就什么也做不了)
    - 什么模型是生成式模型呢？我们想判断Y到底是属于哪一类，在已知这两个分布之后，对K类中的每一类我们都能算出一个概率，那个概率最大，我们就认为属于哪一类

      如果已知 $x|Y=k \sim N(\mu_k, \Sigma)$ $P(Y=k)$

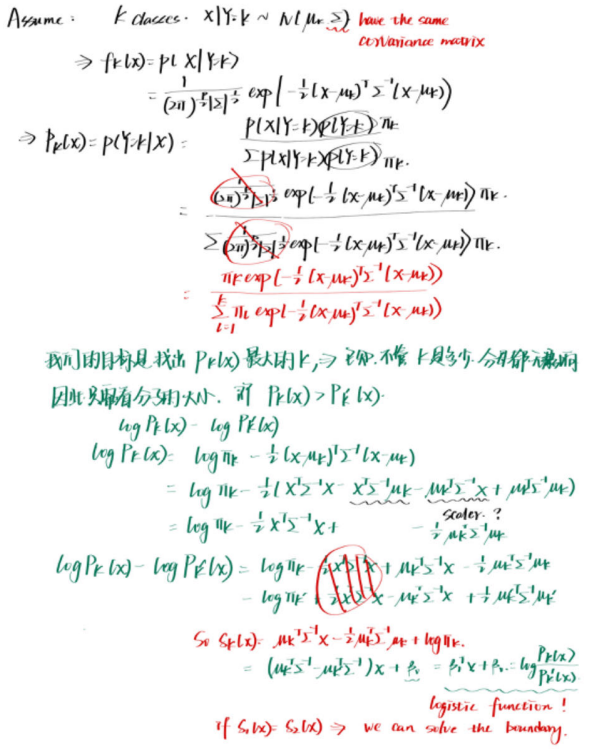
      我们能得到 $y_i \leftarrow P(Y=k)$

      $x_i \leftarrow P(X=m) = \sum_{k=1} P(Y=k) P(X=m|Y=k)$

      $P(Y=k|X) = \frac{P(X, Y=k)}{\sum_k P(X, Y=k)} = \frac{P(X|Y=k)P(Y=k)}{\sum_k P(X|Y=k)P(Y=k)}$

      Then we can get $(x_i, y_i)$

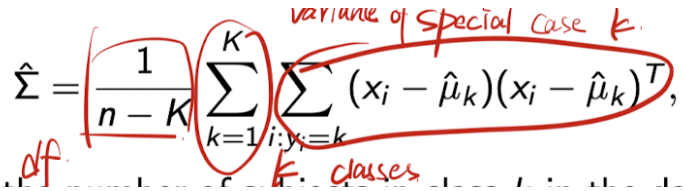
    - 这与判别法不同的在于，判别式模型有一条明显的分界线，在分界线的哪边就属于哪一类，等于是在整个数据集上训练了一个函数，而生成式模型是对每一类都训练了一个函数，即每一类都有一个概率分布
  - LDA theory
    - theory: $\delta_k(x)$ is called discrimination function



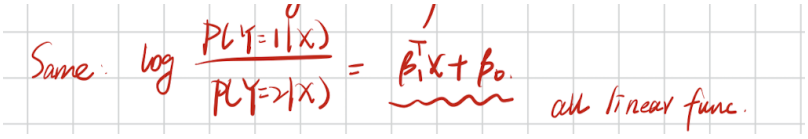

- but how to compute the parameters eg. $\pi_k$, $\mu_k$
  - $\mu_k$ 就是这一类的均值

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i;$$

  - $\sigma$ 怎么计算，首先我们认为所有的都是一个方差，所以必须用到所有的数据，但每一组的均值有不一样，所以还是要分组求，注意除得是什么，是自由度，因为我们有k个均值，因此我们的自由度是n- K（所以之前的方差为什么用n-1来除也是一样的道理）

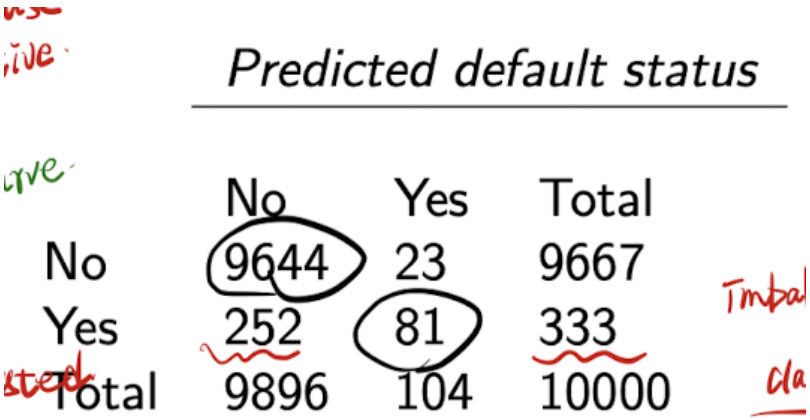



  - $\pi_k$ 就是Y的概率分布，直接用经验分布就好了
  - 还有一个LDA(latent dirichlet allocation): topic model:即给一个document，他可以分类这篇document有40% science， 50% nature and 10% finance类似这种
- Comparison of logistic Reg and LDA
  - same:



- difference:
    - 1. LDA assume the structure so we can just calculate, but logistic Reg don't have the closed form, so we use Newton + MM
    - 2.还记得吗，对于logistic Regression，因为我们用了迭代，当true optimal value is $+\infty$ 时，我们的迭代是不能收敛的，也就是说当两个数据集时完全分开的时候，$\beta$ 必须趋于无穷，概率才能为1，所以是存在numeric problem的，但LDA没有这种问题
    - 3.LDA is a special case of logistic Reg

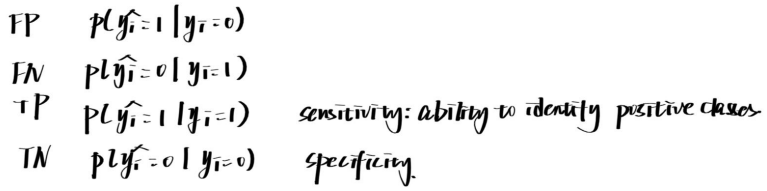
- how to evaluate the method?
    - 很自然的想法是misclassification error, 也被称为averaged misclassification error, or zero - one loss function,但这不是个好的方法，比如下面的例子，如果我们完全不分类，仅仅把所有的都认为是no，我们的正确率会更好，另一种解释是我们更多时候更关注class one，所以我们要换个别的方法来评估



Predicted default status

| | No | Yes | Total |
|---|---|---|---|
| No | 9644 | 23 | 9667 |
| Yes | 252 | 81 | 333 |
| Total | 9896 | 104 | 10000 |

    - 首先我们定义：FP,FN,TP,TN,第一个字母代表我们预测的是否正确，第二个字母代表我们预测的是什么




    - 在每个阈值下，我们都会存在一个sensitivity 和specificity，这个阈值是比如大于0.5还是大于0.2 我们把它归类为1，随着阈值变小，TP和FP都应该升高，因为归类为positive的变多，也就是说，对于一个model，我们有这样的一个随着阈值变化的图，我们称为ROC curve， 这条曲线下面的面积成为AUC，我们把这个数值认为是衡量该模型表现的一个指标
    - 我们还有一种比较model好坏的，去比较loss function，两者的不同是我们这个AUC是没有假设任何模型的，但logistic是假设了一些分布，比如伯努利

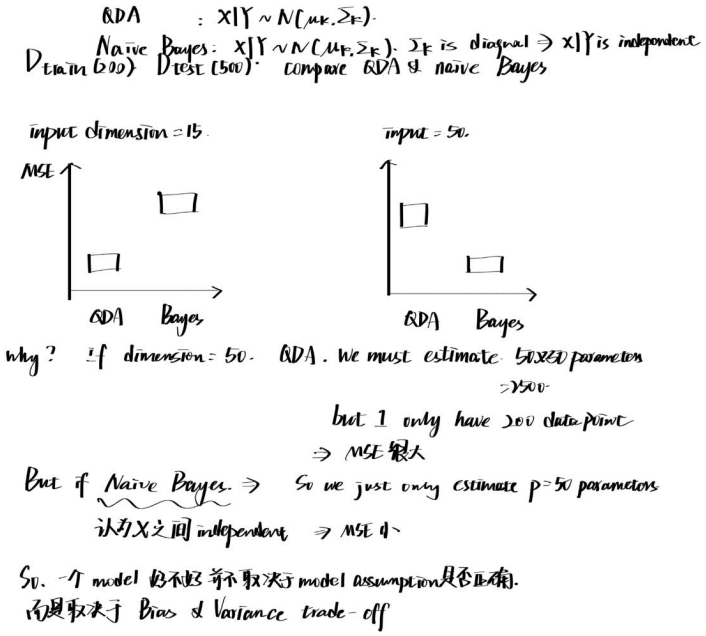
- QDE

- 我们上面的LDA是假设了是线性，也就是大家的方差是一样的，但还有另外一种，我嫩不认为他是一样的，但这样会有很多项就不能消掉了，

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log(|\Sigma_k|) + \log\pi_k$$
$$= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log(|\Sigma_k|) + \log\pi_k$$

This is a *quadratic funciton* of $x$.

- 注意：如果variances 相近，LDA更好，不想尽，QDA更好？！！！！！NO，这个说法是错误的
  - 好不好取决于MSE，也就是biase 和variance，并不取决于assumptions




  - 但是QDA需要更多的数据，如果数据太少，那个LDA可能更合适