**Linear Regression**

Summary

● Simple linear regression

○ Modeling E(Y|X)

○ Assumption

○ Coefficient estimation

* Least square estimation
* Maximum likelihood estimation
* Hypothesis testing

● Multiple linear regression有多个x variable的时候

○ Coefficient estimation

○ Evaluate model performance

○ ANOVA and F test

● **Residual面试常考，单独总结**

○ Residual diagnostics

○ Leverage, standardizing

**Simple Linear Regression**

Classical example

● Let yi be the height of child i; xi be the height of child i’s parent 很可能是正相关的关系

● How to use xi to predict yi ?

● What are we predicting?

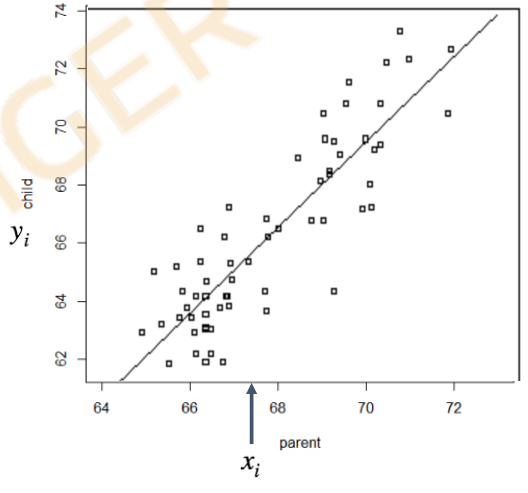
○ **Expected Υ** (可理解为均值) given Χ

● **Simple linear regression**——目的是找出X和Y之间的线性关系，用这条线，Given X to Expect Y

○ Υ = β0 + β1 Χ + ε

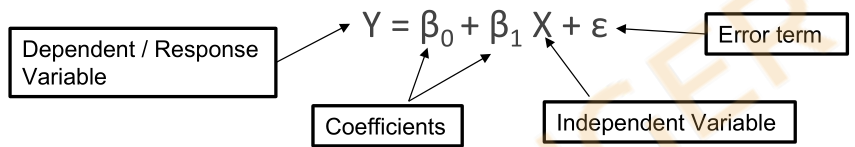
○ Y的期望值Ε[Υ] = β0 + β1Χ

General trend：



一根直线可以被斜率β1和截距β0唯一定义。

**Simple linear regression model**



X——covariance, factors, independent variable.X是已知的，是观测到的值。独立的，不依赖于模型，不随着模型变化的。

Y——dependent variable，Y依赖于X的变化而变化。或Response variable

β0——coefficient

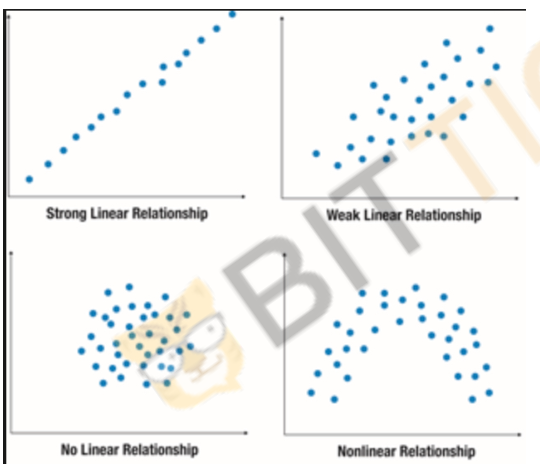
β1——cofficient

ε——error term 每个sample根据生长环境的差异，不会exactly一样，所以有error term。

什么情况下可以build simple linear regression model?

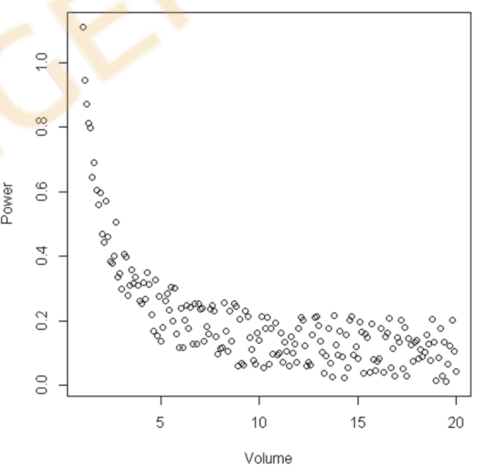
**● Assumptions**

○ 1. **Linear relationship**: Υ = β0 + β1 Χ + ε—— X和Y要有线性关系



图三和图四的区别——图三是X和Y之间没有线性的相关性，是一堆散点，不管X取什么值Y都可能取任何值，没有规律和pattern可循。

图四X和Y之间有相关性，follow a pattern，是弧线的关系，是**非线性的relationship**。（它们之间还是有规律的，有relationship）



Non-linear relationship

○ 2. ε is **independent, identically distributed** (**i.i.d**)独立同分布

相互之间不相关——不同个体之间的 ε互相独立。比如老王家的小孩和老李家的小孩的身高比平均值高还是低多少（和预测值的误差）互不影响，互相独立；

服从同样的分布——爸爸是170，小孩有比较高的概率168-172，有比较小的概率160或180。分布一样，都不会偏移太多。

○ 3. **Homoscedasticity** (**constant variance**) Var[ε|Χ = x] = σ^2 (no matter what x is)

ε会有一个浮动的范围，有variance，但这个variance和X的值无关。不管X取什么值，ε的variance不变。

以上，前三点是必备的assumption。一定要有。

以下，第四点，not necessary in your model但通常要有。（理论上Not necessary,但实践中都会加上。面试如果问到还是要把第四条说上。）如果模型符合这一点，model的很对推论、推导will be much easier,也会有很多很好的性质。所以通常还是会加上这一条assumption。

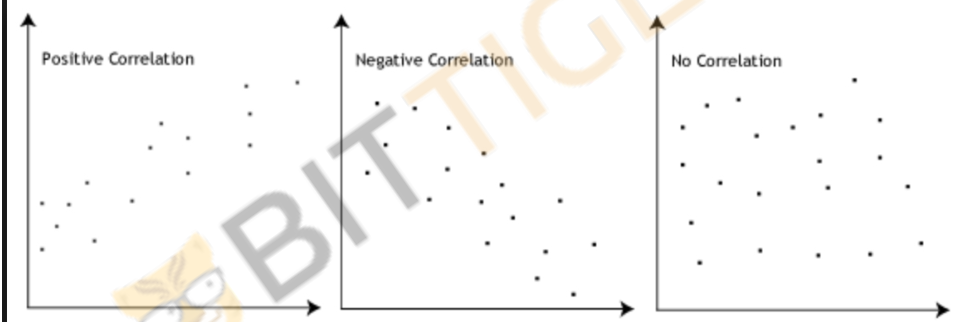
○ 4. Gaussian Noise: Normal distribution ε ~ Ν(0, σ^2 )

希望 error term ε符合正态分布。比如180的爸爸生的小孩的身高差基本符合正态分布。（大部分人差别比较小（集中在中间），小部分人差别特别大（落在两边））

* 举例说明违反assumption的一些例子： **Assumption Violation**

怎么样的就算不符合simple linear regression 的assumptions?

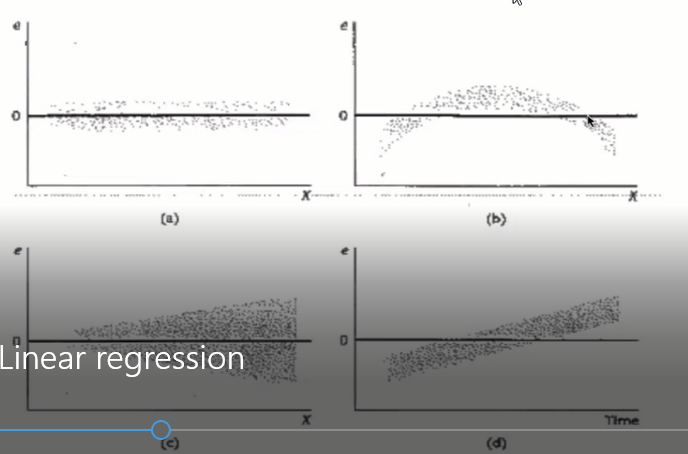
1. no linear relationship / non linear relationship
2. ε Dependent on each other



Y轴是ε-error term，X轴是第几个observation

第一个图说明ε和观测顺序正相关，第二个图说明ε和观测顺序负相关。第三个图是不管你什么时候观测到的，ε比较随机。

1. Heteroscedasticity (non-constant variance)



横轴是第几个X，纵轴是error term。 Constant variance of error term就是指variance独立于X，不论X变大变小，variance不太变。看是否variance constant就是垂直拉一条线，和点区域重合部分的**宽度不变就是variance 不变**。

（只有）a是homoscedasticity,随着X的变化ε的分布不变（包括中心值(mean)不变+宽度(variance)不变）。

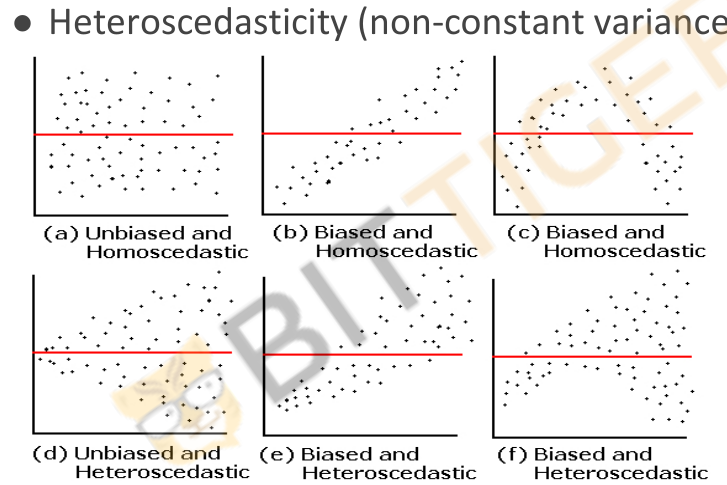
我们要找homoscedasticity，就是要找长的像a这样的图。ε随着X的变化，它的范围不怎么变化，均匀地在0上下浮动。

b和d也是constant variance（宽度相同）但是~~不能叫homoscedasticity因为~~不同的X对应的ε的分布不同。a的ε的分布就是相同的，不同的X对应的ε都是以0为中心上下浮动差不多的范围。b的中心随着X的变化而变化即均值一直变化所以符合的distribution也一直变化。d的横坐标是Time，ε和时间有相关性了，随着时间变化，也不能叫homoscedasticity.

c的variance（ε的浮动范围）随着X的增大而增大。不是constant variance。

我觉得这里老师把assumption23结合起来了，独立同分布+同variance是她这里的“homoscedasticity。”（所以homoscedasticity到底光是constant variance还是还要加上同分布？？？？）不然就跟下面矛盾了。我的理解是只有a满足assumptions，bcd都violate了assumptions，**要看是否满足simple linear regression 的assumptions找a那样的ε。bcd不行**，就可以了。





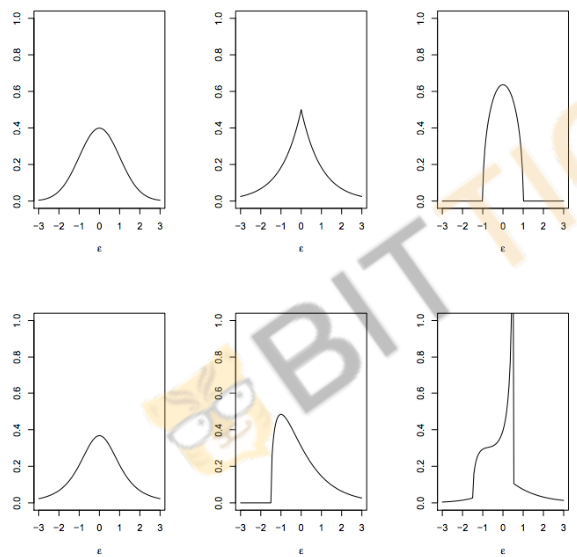
**如果点的取值范围在一条横线上下浮动，即均值/中心值不变，就叫unbiased**。Like a and d.

如果均值改变了就叫biased。Like b c e f。

abc的宽度不变即variance不变还是叫它们Homoscedastic.

def的宽度变化了即variance变了就叫它们heteroscedastic。

1. Some error term example——除了Gaussian noise, error term还能长啥样？还有什么其他的选择？



第一个图是Gaussian noise，正态的，normal distribution的noise。

为什么prefer ε ~ Ν(0, σ^2 )而不想选其他的奇形怪状？Why Gaussian noise?

因为当我们有正态分布的时候，会有很好的性质。比如CLT近似出来的那些。（当样本量足够大时可以用正态分布的均值来近似population mean）

● Central limit theorem

○ Noise might be sum of lots of little random noises from different sources, independent and with similar magnitude.

● Mathematical convenience

○ Closed form estimation 数学推导变得much easier。

**Under** **Gaussian Assumption**

Υ = β 0 + β 1 Χ + ε

解释：为什么在Gaussian Assumption下，E(Y)会把ε去掉? E(Υ) = β 0 + β 1 Χ

**β0, β1和Χ都是常数**，在建好一个model时这些都是已知的。E(ε) = 0. 对Y取期望值会把error term去掉。

ε = Υ - β 0 - β 1 Χ ~ Normal (0, σ^2 ) ->

**Y = β 0 + β 1 Χ + ε ~ Normal (β0 + β 1 Χ, σ^2 )**

**Both Error term and Y(given X) follow Normal distribution.**WHY?

Error term follows normal distribution is our assumption.

(For given X,) Y follows normal distribution because

**If X~N(miu,σ^2), then X+a ~ N(miu+a,σ^2)----assume a is a constant**

ε ~ Normal (0, σ^2 ) -> Y=ε+**β0+β1Χ** ~ Normal (β0+β1Χ, σ^2 )

到这里，我们知道了gaussian assumption下error term和Y的分布。 为什么要知道这两个分布？

**Coefficient estimation**

在Linear regression的建模过程中，最重要的是estimate β。Y和X之间存在真实的关系， β0和β1都是真实存在的，但你不知道它们的真实值是多少，需要用许多数据去estimate。

● Some definition

○ Sale price Y: dependent variable, output, response

○ Lot area X: predictor, independent variable, covariate, input

● **Coefficient estimation**

○ Remember assumption Υ = β0 + β1Χ + ε

○ How do we **find optimal (β0 , β1)**?如何找到最有可能接近真实情况的β。两种方法。面试不用推导但要知道概念。

Least square error estimation (LSE)

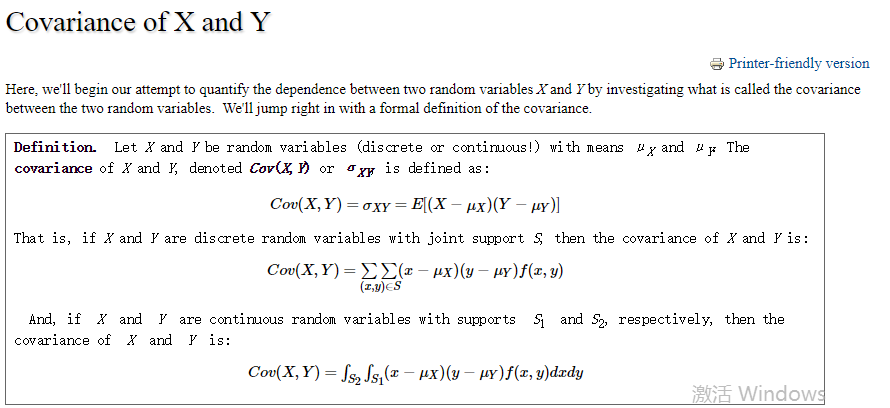
Maximum likelihood estimation (MLE)

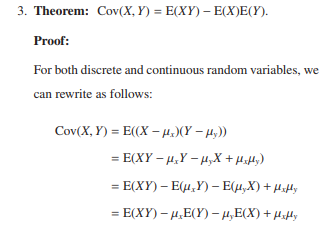
Compare LSE and MLE

LSE

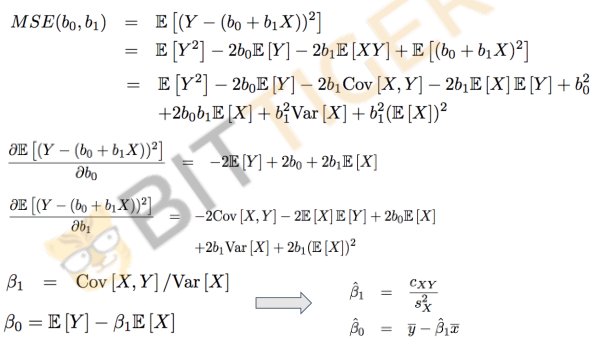
Least square error estimation——想让model估计的Y=b0+b1x和真实的Y差距（即error）越小越好，误差越小模型越精确。不同观测值的error如果直接相加会正负抵消。可以对error求平方来避免正负抵消——squared error。

找出让平方差最小的β\_hat。β\_hat为 β的估计值（我们不知道β的真实值是多少，用这种方法估计β，估计值称作 β\_hat）

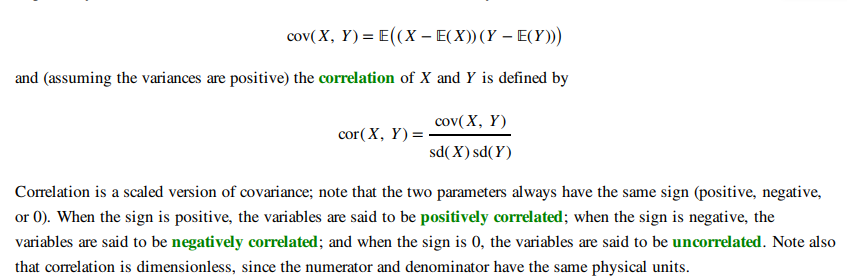


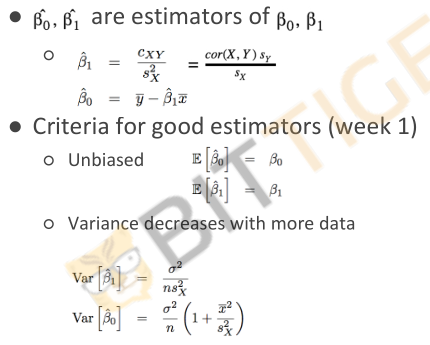


● Mean square error (MSE)



求 使MSE最小的β——MSE对β0求导等于0，MSE对β1求导等于0-->两个等式，求出β0和β1.





Unbiased——每次对β0的预测值可能不太一样，（真实值是固定的，但估计值肯定有变数，随着每次data不一样，每次预测出的β可能不太一样）和β0的真实值有偏差，但我们**希望β0的观测值的期望**（多个β0观测值的平均）**= β0的真实值**。对β1同样适用。

Variance——每次对β0的预测值可能不太一样，我们**希望β的观测值的variance越小越好**。当数据越多的时候，信息越多，预测的不确定性越小，这个variance越小。**Sample size越大**，coefficients estimation 的**不确定性(variance)越小。**

经过数学家证明，通过以上Least square error estimation计算得出的β符合这两个性质（Unbiased and variance decreases with more data）。

LSE是从“观测值和预测值的差越小越好”的角度来estimate coefficient β.下面的算法MLE是从概率的角度计算β。

MLE

What is likelihood

● Probability

○ If we know β0 = 10, β 1 = 0.8, E(Y) = 10 + 0.8X, possibility to observe (x1 =70, y1 =72), …, (xn,yn )?

● Reverse above logic

○ **If we observe (x1=70, y1 =72), …, (xn , yn )**

**○ Likelihood (L) of β0 = 10, β1 = 0.8?**

● Connection

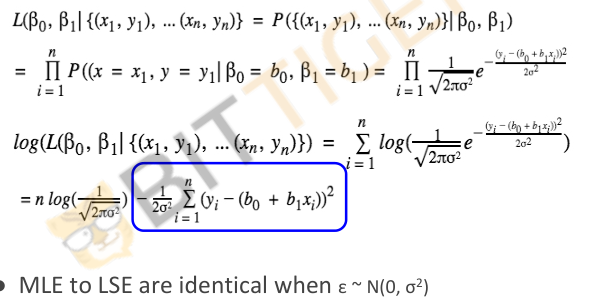
○ L(β0,β1|{X, Y}) = P({X,Y}|β0,β1)

● How to use likelihood to estimate β0,β1

○ **Maximizes Likelihood function L(β0,β1|{X, Y})**

Maximum likelihood estimation (MLE)

● **Under Gaussian error distribution**



Given X----估计值Y随着β的变化而变化，如何找到β\_hat，使得给定X，真能观测出估计值Y的likelihood最大？

**总结区别**：

LSE是最小化观测值和预测值之间的平方差，是通过减少误差的方式计算出的coefficients的estimation。

MLE是通过最大化 能使给定的那些观测值出现的β 的可能性 来估计系数的。

这两个方法在高斯假设下殊途同归。当ε ~ Normal (0, σ^2 )时，这两种方法结果一样，只是计算过程不同。当不符合正态分布这个assumption时，这两个estimater的计算结果不一样。（所以在实践中，为了简便，希望条件尽可能满足高斯假设。）

面试中这里怎么问？——你怎么estimate coefficients？

回答——可以用LSE（least square error estimater）或MLE(maximum likelihood estimater)，这两种都可以。

问这两种有什么区别？

回答——LSE 是通过最小化squared error, MSE是最大化likelihood.它们两个在ε ~ Normal (0, σ^2 )的情况下是一样的。

**Hypothesis testing for coefficient**

**——用来检测Y和X到底有没有关系**

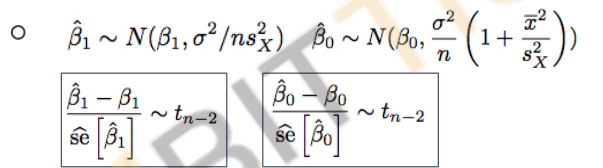
我们想知道β是否为0。比如β1 = 0说明X和Y之间没有线性关系。此时Y = β0+ε。

检测β是否为0——用T-test。

● Test assumption

○ Null hypothesis β1 = 0 (What type of test)

● Calculate stats: compare observed and H0



（我理解的两个~左边都是在对β0\_hat和β1\_hat做标准化（减去均值除以standard error），如果var已知，标准化后的结果应该符合标准正态分布N(0,1),这里var未知，所以符合t-distribution。）

○ Why t distribution. Recall t score. How to decide d.f. (degree of freedom)?

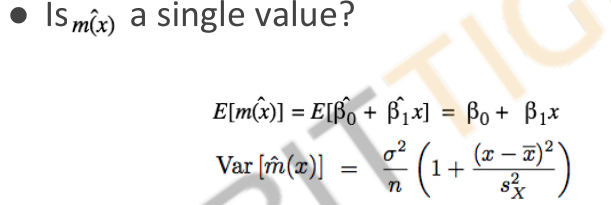
因为β\_hat0,1都服从正态分布，**β\_hat的variance都是未知的**，（这里我的理解是sigma虽然是常数，但未知，不知道error的variance是多少，所以β0和β1的variance也未知），所以它们就服从t-distribution。

**How to predict for new data?**

● New data point x, how to predict y?

● Predicted (fitted) value at x:

m\_hat(x)可以理解为y\_hat。



预测出的值有不确定性（variance）

○ **Variance grows as σ^2 is larger**; **more noise in predictions 误差项的方差越大，预测结果中的噪声也越大。**

○ **Larger n is, smaller variance is; more precise of predictions 样本数量越大预测值越精确，（不确定性越小）**

**Distribution of predicted value**

● **Under Gaussian error assumption,**  **follows normal distribution**

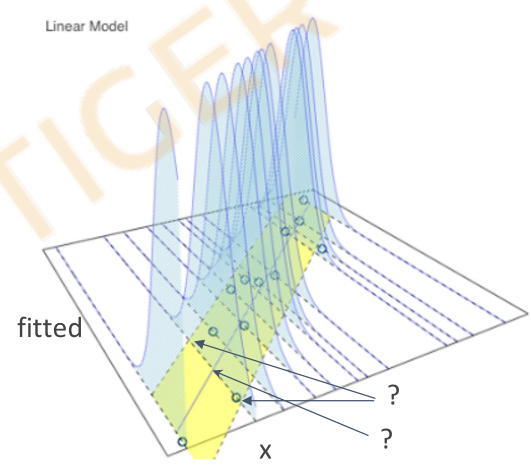
○ What’s Expectation-E(Y) = β0+β1X,图中底面实线。 Variance, confidence interval?

● Visualization

○ Observations as circle

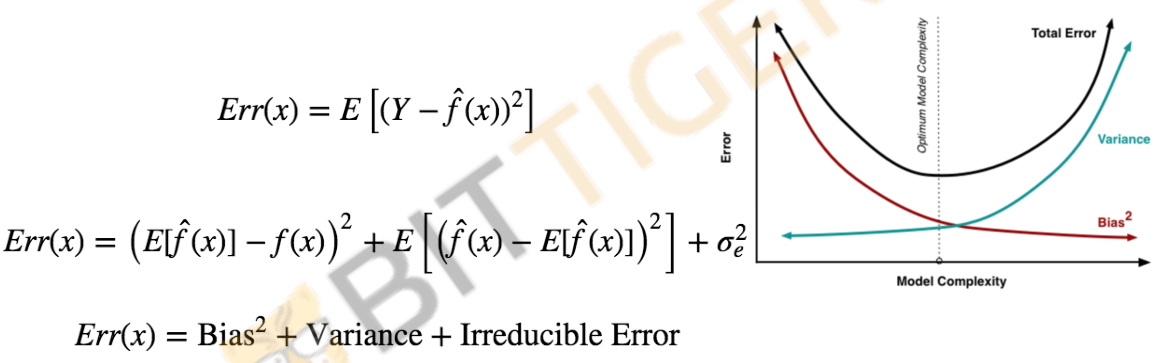
○ Mean prediction as solid line

○ [5th, 95th] quantile of predicted value.



**Bias and variance tradeoff**

● Prediction error, suppose we have Υ = f(Χ) + ε



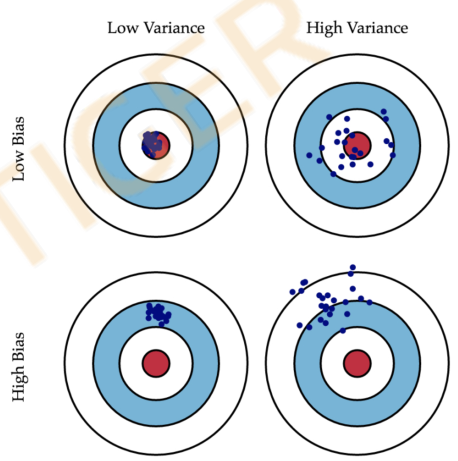
Bias=预测值的期望-真实值。E[f\_hat(x)] = E[β0\_hat+β1\_hatX] = β0+β1X = 预测值的期望；f(x) = 真实值。

Variance(f\_hat(x)) = E[(f\_hat(x) - E[f\_hat(x)])^2]

Bias and variance

● Accurate (small bias)准确性

● Precise (small variance)确定性



蓝点是模型预测值，红点是真实值们，BIAS看的是离红心的距离。蓝点离红心越近说明预测的准确度越高-Low bias

Variance看的是蓝点的离散程度。越分散，不确定性更高（到处落）；越集中，不确定性越小，low variance。

因为bias和variance的trade off,很难有1,4的情况，通常是图1图3.（/对角线）在右上角和左下角之间取舍。

**Multiple Linear Regression**

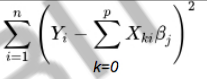
**——有More than one predictor-Xi**

● More than one predictor, say p. ith data point Yi-第i个data point的真实值。

Yi = β0+β1X1i+β2X2i+…+βpXpi+εi

所有的推导在原理上都一样，**还是LSE和MLE**。只是从一维变成多维，要用矩阵。

● How to get estimate (LSE)?



○ MSE: 1/n \*

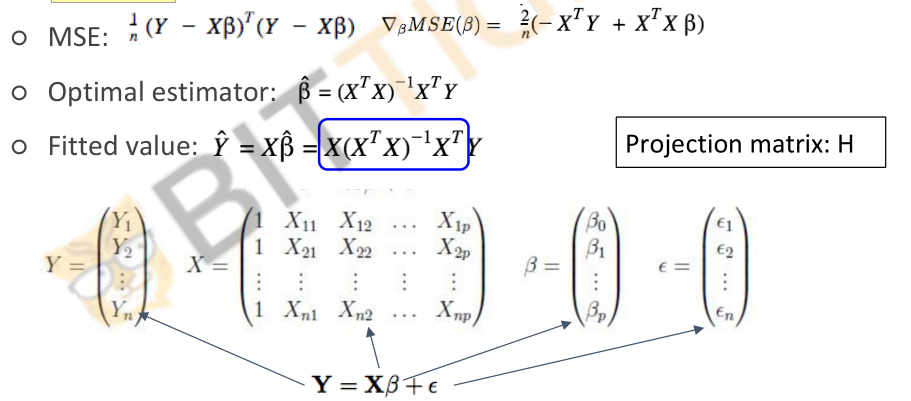
○ (β0 , β1 , … βp ) minimizes MSE (derivative = 0).

LSE for multiple regression

● Matrix form

○ n\*1 matrix Y, n\*(p+1) matrix X, (p+1)\*1 matrix β , n\*1 matrix ε

Υ = Χ β + ε



How do you know if your model works?——面试中经常问

● For one covariate——一维，simple linear regression.测试β1是否为0.

○ T test for βi

Null hypothesis: β i= 0

● For multiple covariates

如果想**单独看每个**feature X对应的**βi是否为0，**还是用**T-test**

如果你想**看所有的β**(β1到βp)**是否都等于0**，这时要用**F-test**。

○ F test for β1, β2,...,βp

Null hypothesis:β1= β2=...= βp = 0



**ANOVA方差分析**

● ANOVA (Analysis of variance)——有很多用途。这里只讲用于regression model中的用途。

1000个数据点，1000个Y值互不相同，所以会有个variance——观测值的variance。我们用ANOVA把这个观测值的variance分解到不同来源。第一种来源——regression model。Model能解释一部分为什么Y之间有差异。（计算时用Model的预测值减去观测值即真实值的平均值。）还有一种Y之间的差异是model所不能解释的。另一种来源——residual，即error term造成了剩余的variance。（就算我知道父亲身高和儿子身高之间的关系的model，还有一部分个体差异是model无法解释的，由于其他因素的影响比如小孩的生长环境，营养状况，活动量等造成的variance）。

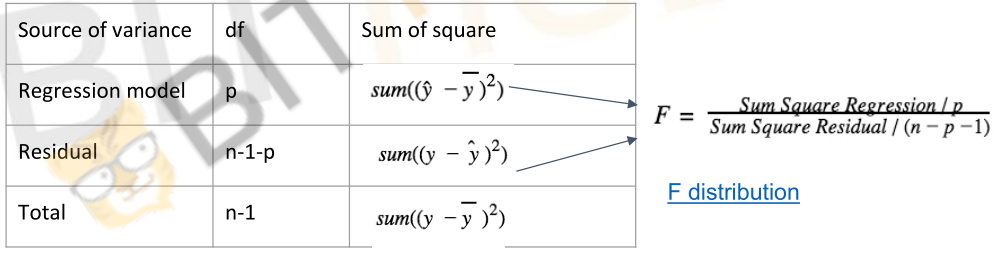
**total variance**（真实值-真实值[即观测值]的平均数）**= model解释的variance**（预测值-真实值[即观测值]的平均数）**+ residual variance**（真实值-预测值）**——将variance分解为了两个部分——ANOVA在干什么**

○ **Decompose variance into sources**

○ F test

Null hypothesis:β1= β2=...= βp = 0

Calculation: (p does not include β0 )



df——degree of freedom

Model的df是p，因为有p个X（feature）在变化。

Total df是n-1，因为有n个样本。（我的理解是你可以取这N个点作为样本，也可以取另外N个点作为样本，这N个点的取值是自由的）

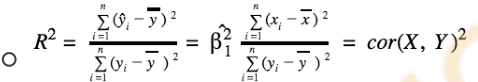
那么residual的df就是剩下的，n-1-p。

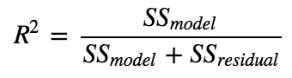
**F-test就是看能解释的部分和不能解释的部分的比例。**

**How to evaluate model performance?**

**记住！**

● R^2 , and relation with correlation





● Property

○ **R^2 , percentage of variance explained by model**

**模型能解释多少数据上的差异。**

**建模的目的是为了预测Y——希望模型能解释的Y的差异越多越好**

即我们希望模型解释的variance占总variance越大越好

○ 0 <= R^2 <= 1

○ R^2 could be misleading

**More features always increase R^2**

Need to adjust R^2 

变量多的话，**加上一个penalty**。**Adjusted R^2会随着p的增大而减小。**

第一个model-两个变量解释了50%的variance，第二个model-10个变量仅仅解释了60%的variance。第二个model的效率低。

**Adjusted R^2中加入了# of features**-p,可以理解为**看的是model performance的平均绩效**。

这部分需要记住的是：**一般来说，要看一个model的performance，看的是R^2，但有时R^2会misleading（会鼓励你尽可能多的加feature，但feature太多其实并不好，增加模型复杂度等等问题，还是尽量用比较少的feature达到差不多的目的即可），所以要看adjusted R^2（对增多feature会有penalty）。**

**面试可能会问：What is R^2? 回答：percentage of variance explained by model即可。**

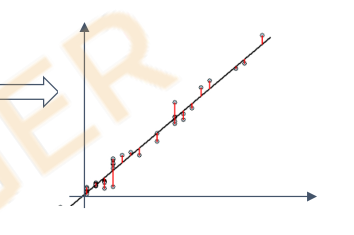
**How do you know your model meet assumptions? Residual Diagnostics余差分许**

**非常重要，面试时非常喜欢问**

**怎么知道你的模型符合假设，或者你怎么知道你选的Linear regression模型是对的？**

**Residual vs. Error Term**

****

● Residual y的观测值-y的模型估计值

Observed value - the estimated value of the population mean



● Error (Noise)

y的观测值-y的真实关系的估计值(population mean) =真实的noise

Observed value - Unobserved true value of population mean

○ residual ei is an observation of error term εi

residual is an observable estimate of the unobservable statistical error.

Error term是一个随机变量，不同的Sample和sample之间error term

的观测值（即residual）都会有不同。

Error term - disturbance

Residual-fitting deviation

One can standardize statistical errors (especially of a normal distribution) in a z-score (or "standard score"), and standardize residuals in a t-statistic, or more generally studentized residuals.

● Properties of residual

○ 

○ Constant variance, unchanging with x.

○ The residuals are not uncorrelated with each other, or extremely weak

○ Normal distributed

————————————————————————————————————————————————————

**下面是一些画图方式，用以检查是否符合linear regression的assumptions**

**Residual v.s. Fitted value**

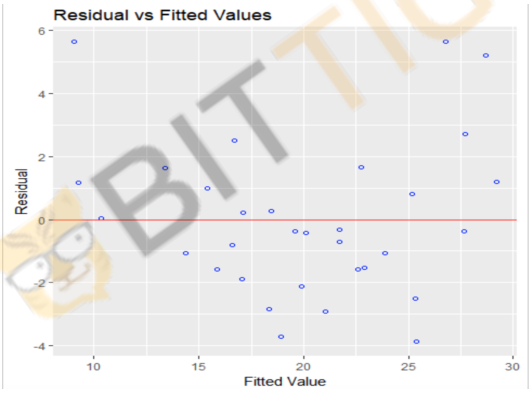
画一个纵坐标是residual，横坐标是y\_hat的图，来检查结果是否unbiased（以零为中心）和是否有constant variance。

● To check if unbiased and homoscedasticity

• The residuals spread randomly around the 0 line indicating that the **relationship is linear**.

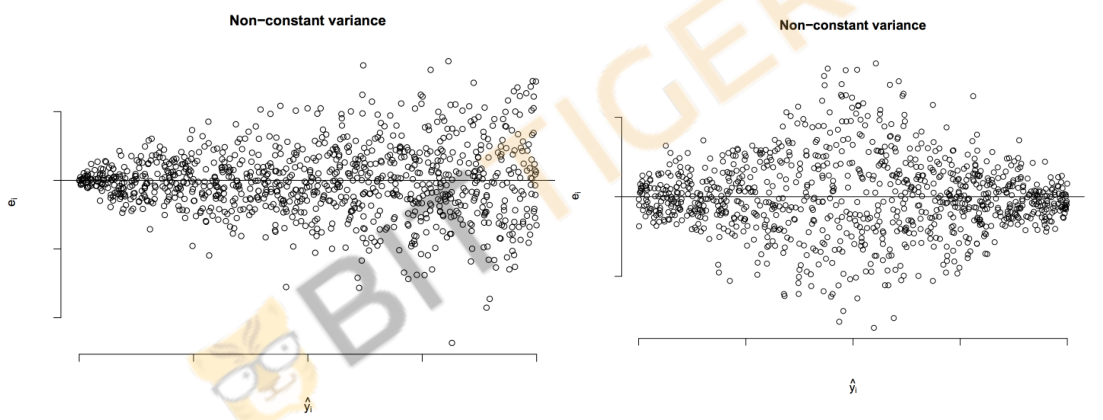
• The residuals form an approximate horizontal band(宽度不变)around the 0 line indicating **homogeneity of error variance**.

• No one residual is visibly away from the random pattern of the residuals indicating that there are **no outliers**.



这个图的residual并没有spread randomly around the 0 line，两边的residual在零上，中间的residual在零下。整体呈碗形，和X轴有关系的变化。 宽度也不太一样。 violate了assumptions。

Residual v.s. Fitted value – Violation Examples



**QQ Plot**

R中可以直接画这个图

横坐标画Theoretical quantile,是理论上normal distribution的quantiles；纵坐标画sample quantile，是数据中的的quantiles。比较这两个来源的quantiles，看它们是否基本match（1:1，基本沿着对角线分布），来检查sample是否符合正态分布。

● To check for Normality

• The points should roughly follow the diagonal straight line

图一不满足normal distribution，图一中的红线不是对角线，并且右上角还有几个点跑远了。

图二属于比较好的QQ-plot，两边可能有一些不太符合的，但不会差太远。（也没有跑的很远的点）

图二是从N(0,1)拿到的sample，但却不完全符合——因为毕竟是样本，和理论值还是有差异，而且样本数据量不大时更不可能完全符合。样本数据量越大就会越符合。

•You can also **check normality use Shapiro-Wilk test or Kolmogorov-Smirnov test**

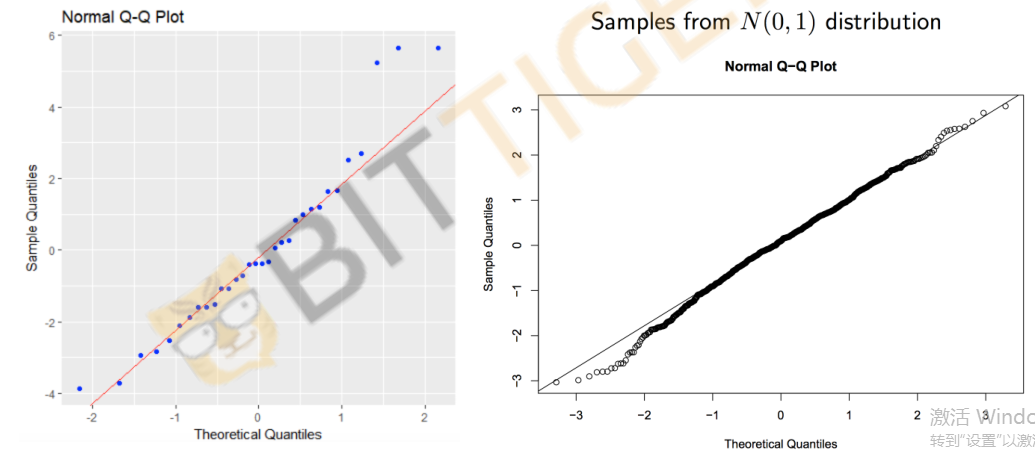
面试经常问：什么情况下MLE和LSE相等？

答，符合normal distribution assumption的情况下。

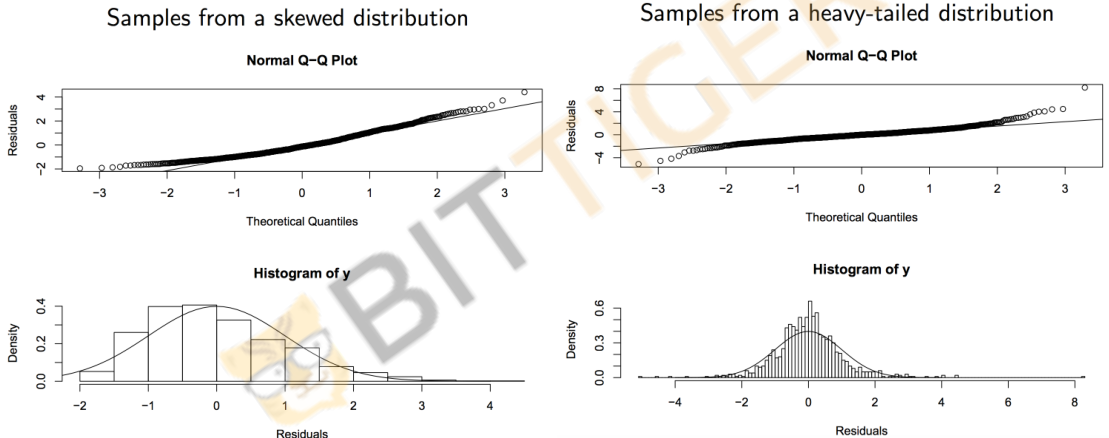
怎么检查符不符合？

可以画QQ-plot（大多数情况下都用QQ-plot来检查normality，因为根据CLT，在data足够多的时候近似正态分布，基本上不会有太大的问题，QQ-plot不要差太多就可以了），或者run S-W test/ K-S test这样的normality test（这两个test有点过于严格，很难pass这个test所以一般不跑，用QQ就够了。除非特别boarder line，根据QQ-plot觉得不太正态，但也不太确定，可能就run一下这个test）。

其实很多时候就是看QQ-plot都不正态，我们知道它不正态，但也继续跑了这个model，因为model perfectly符合所有assumption的情况很罕见，很多时候知道有一些问题还是会去建这个model。



QQ Plot – Violation Examples

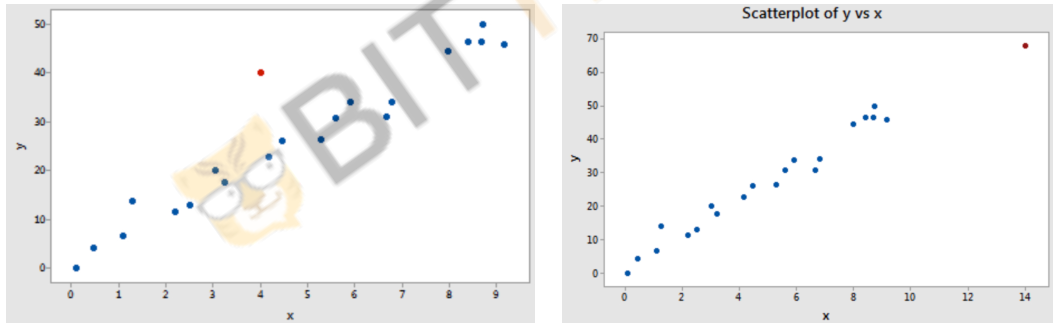


**Outlier, High Leverage, Influential**

• An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.（x在正常范围但Y不正常，表现为该点不在其他data的general trend上）

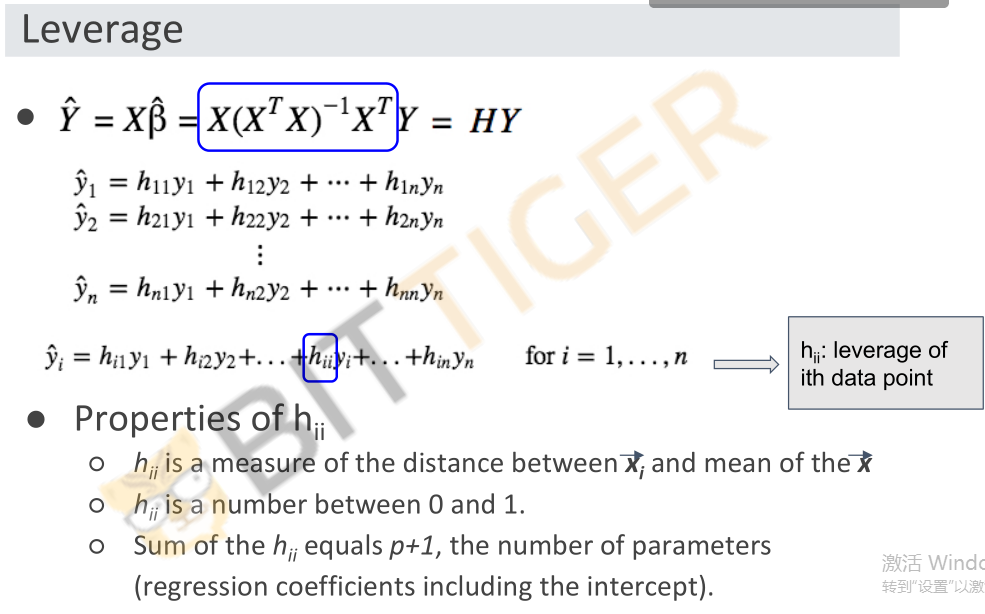
• A data point has **high leverage** if it has "extreme" predictor x values. With or without this data point will highly impact the estimate of coefficients（x不在正常范围内）

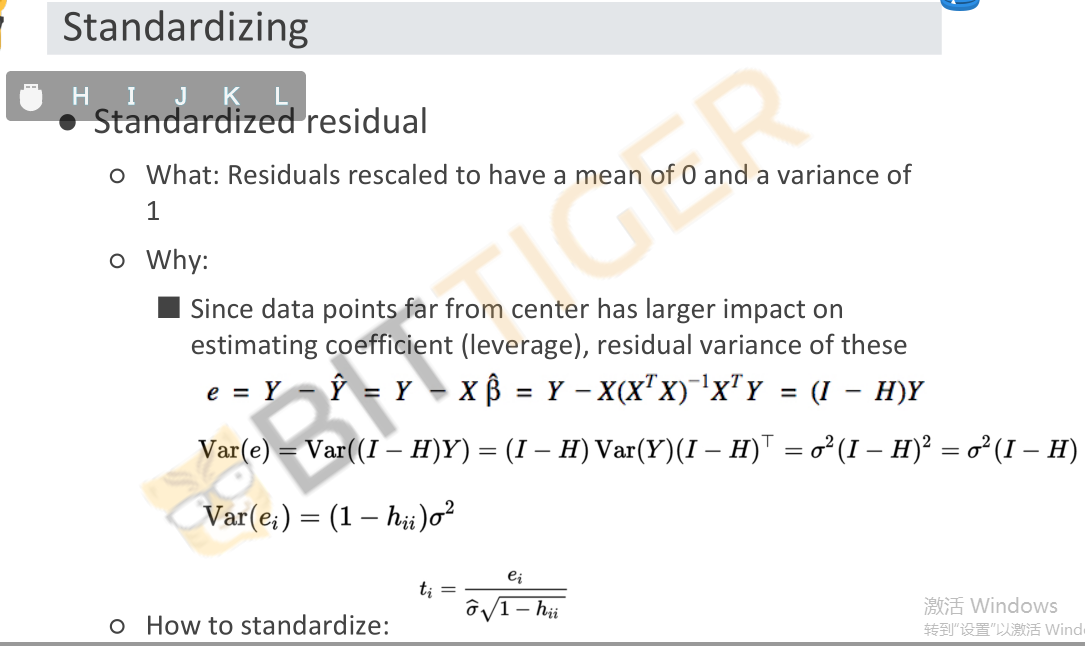
• A data point is **influential** if it unduly不适当地，过分地influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. **Outliers and high leverage data points have the potential to be influential**, but we generally have to investigate further to determine whether or not they are actually influential.

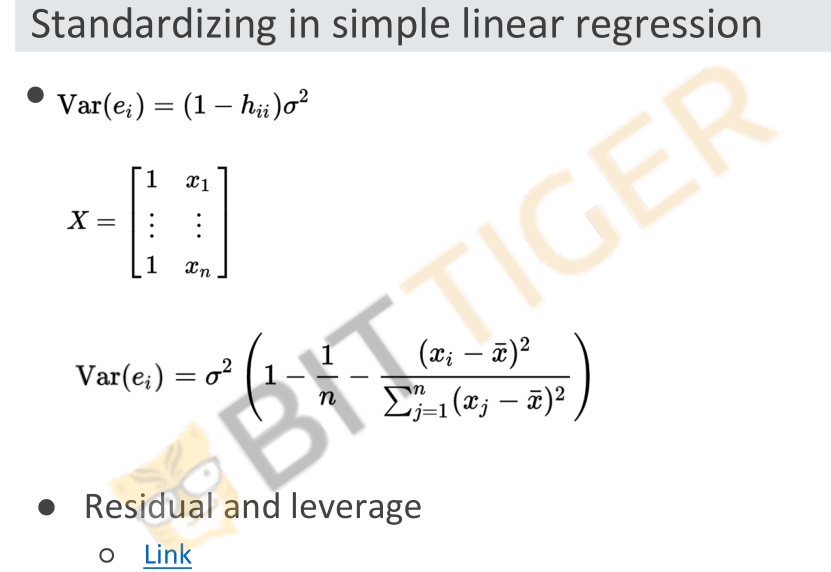


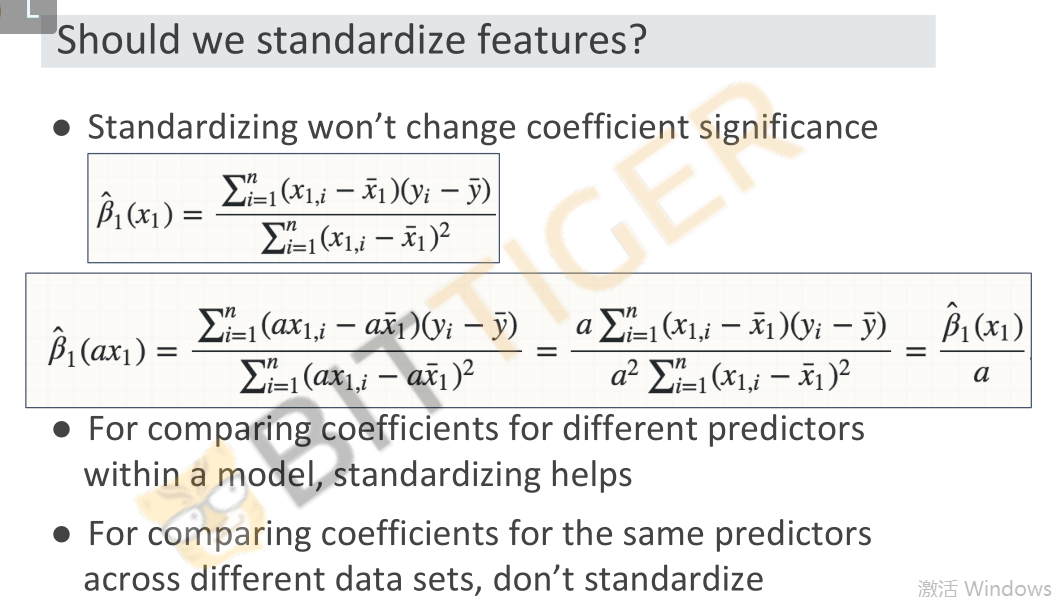
<https://newonlinecourses.science.psu.edu/stat501/node/337/>

整套课程很有用，偏实践。









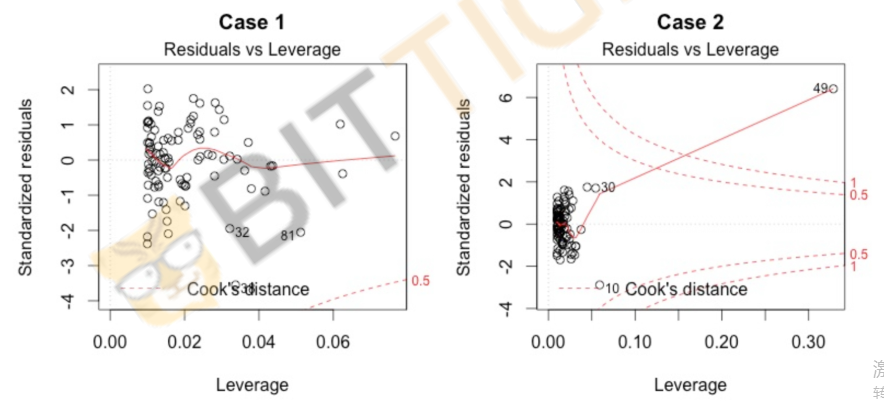
**Standardized residuals vs. Leverage**

● To check high leverage, influential data points

R中可以直接画这个图

• Leverage: consider fitted line as lever杠杆, passing center point. Points further from center point have larger leverage. 横坐标是leverage，越往右，leverage越高。h\_ii = [H]\_ii measure distance of x to center of x。但不是说leverage高就都有问题。图中红色的虚线叫做cook’s distance，用于判断这些点是否influential。Outlier和high leverage都有可能出现在cook’s distance以外：

• Rule of thumb: Cook’s distance>1 -->influential points，该点对模型会有很明显的影响。



可见，左图没什么influential points，右图的49号data point是个high influential data point。

**面试经常问：我有一个high leverage/influential point，该怎么办？如何处理这些点？**

**How to Deal with Problematic Data?**

Do NOT simply delete them by default!!!

You must have a good, objective reason for deleting data points

1. Check for obvious data errors.比如-1999年，身高179米。correct or delete them

2. Consider the possibility of mis-formulated model:模型是否不合适

• Did you leave out any important predictors? 漏掉了重要变量以至于产生了outlier。

• Should you consider adding some interaction terms?

• Is there any nonlinearity that needs to be modeled?-generalized linear regression

3. **If you delete any data after you've collected it, justify and describe it in your reports.**

If you are not sure what to do about a data point, **analyze the data twice** — once **with** and once **without** the data point - and report the results of both analyses.“我不确定这部分data是什么情况，我这里有两个结果”

具体处理，来自网页psu-STAT501

Copied from <https://newonlinecourses.science.psu.edu/stat501/lesson/11/11.7:>

11.7 - A Strategy for Dealing with Problematic Data Points

You should certainly have a good idea now that identifying and handling outliers and influential data points is a "wishy-washy" business. That is, the various measures that we have learned in this lesson can lead to different conclusions about the extremity of a particular data point. It is for this reason that data analysts should use the measures described herein only as a way of screening their data set for potentially influential data points. With this in mind, here are the recommended strategies for dealing with problematic data points:

First, check for obvious data errors:

If the error is just a data entry or data collection error, correct it.

If the data point is not representative of the intended study population, delete it.

If the data point is a procedural error and invalidates the measurement, delete it.

Consider the possibility that you might have just misformulated your regression model:

Did you leave out any important predictors?

Should you consider adding some interaction terms?

Is there any nonlinearity that needs to be modeled?

If nonlinearity is an issue, one possibility is to just reduce the scope of your model. If you do reduce the scope of your model, you should be sure to report it, so that readers do not misuse your model.

Decide whether or not deleting data points is warranted:

Do not delete data points just because they do not fit your preconceived regression model.

You must have a good, objective reason for deleting data points.

If you delete any data after you've collected it, justify and describe it in your reports.

If you are not sure what to do about a data point, analyze the data twice — once with and once without the data point — and report the results of both analyses.

First, foremost, and finally — it's okay to use your common sense and knowledge about the situation.

**常见的、关于linear regression的interview questions：**

Basic

• What model will you use? For **continuous** Y variable, linear relationship，就可以考虑用Linear regression.

（for categorical y variable，linear regression不合适）。Y必须是连续变量，因为要符合normal distribution。

• What are the assumptions of linear regression? How to check?

见以上笔记。

• How is coefficients estimated?

LSE和MLE，they are identical under normal distriburion当ε ~ Normal (0, σ^2 )时.

• What kind of diagnostics will you do?

1. test for bi and F-test for b1=b2 =... = bp = 0——Hypothesis testing

R^2 and adjusted R^2 to see the percentage of variance explained by model——Model performance

Residual diagnostics to see whether model meets assumtions and influential points.——Standardized residuals vs. Leverage

Cook’s distance > 1?

Advanced（面试比较难的职位会考）

• What is the estimate of coefficients?可能会让你写出推导 How to derive it?

• What is variance of beta, error term? Degree of freedom这些东西怎么算

• What to do with non-constant variance? Non-linear? Non-normal?——generalized linear regression中的内容

<https://newonlinecourses.science.psu.edu/stat501/lesson/15/15.5>

• How to improve model performance?

1, take home project report

2, product sense questions (2 hour, part of it)

3, industry what function, dynamic pricing, marketing analytics, growth, risk

4, What BA do, how to grow to a DS / work experience, day to day work

5, Q&A

答疑：test data和training data要在一个范围内，有可能training data的范围内是这个model，当超出这个范围就不是这个relation了，不好预测训练数据范围外的数据点的response。

Summary

● Steps to build a model

○ Problem statement

○ Feature processing

○ Feature engineering

○ Features selection

○ Model evaluation

● Build linear regression

○ Coefficient estimation

○ Residual variance, p value, F test

○ Residual diagnostics

○ Model performance