---------- 1.
http://www.mitbbs.com/article/Statistics/31357771_3.html
Q: If each of the two coefficient estimates in a regression model is statistically significant, do you expect the test of both together is still significant

我觉得是不是应该从两个variable是不是correlated(high or perfect multicollinearity)的角度讨论。如果是highly correlated, 那么remove one variable does not lose explanatory power, 所以用F-test测试 b1=b2=0的时候，应该是significant的，因为两个b至少有一个应该不是0，我们就应该reject H0, conclude significance..

第二种情况是没有correlation, independent variables, 那么也应该是significant的，因为observing two unlikely independent events is "more unlikely" than observing them alone..

所以结论是significant for both cases.. 我说的是不是很扯呀。。

A: 假设y x1 x2都已经中心化和标准化
y=b1*x1+ b2*x2 + e

记T1^2为 y=c1*x1 + e的score statistic （数值上大致等于Wald stat的平方）
T_2^2 为 y=c2*x2 + e的score statistic
简单计算表明H0: b1=b2=0的score statistic有形式

T= (T1^2 - 2*r*T1*T2 + T2^2)/(1-r^2)
r是correlation(x1, x2)

所以对于绝大部分情况，给定T1^ 和 T2^2充分大，且 r < 0时，你的回答应该是对的。如果r > 0，则不一定，因为r对于分子分母的影响是同向的。这也是为什么对于两个负相关的因子，联合起来考虑通常能提高power；对于正相关的情形，则未必，因为检验统计量的增大（甚至减小）未必能抵消df增大的影响。对于独立的情形，r=0，这时候T=T1^2 + T2^2 ~ chisq(df=2)，大概也是对的，要看具体的significant threshold。
so depend on the correlation, because both test statistic and df are changed, it could be sig or non-sig

---------- 2.
http://www.mitbbs.com/article/Statistics/31357929_3.html
下面是面经
1.research experience （讲了15分钟）
2.一堆data, Bernoulli distribution, 40% = 1, 60% = 0, design test to see if

40% is significantly less than expected?
我说用chi-square, 因为是non-parametric, expected frequency = 0.5, 他又问chi-square的公式。
3. 另一堆data, 70% = 0， design a test to see if the two samples （和前面一道题的60%） are different, 我说用two sample t test, 这个对不？
for 2 and 3, use normal approximation, construct Z test for one sample and two sample comparison of proportions
4. 3 columns of data, two continuous variable, one boolean variable, design experiment to see relationship between response and variables
我说用linear regression, 他说要一步步怎么做要说清楚，我说先center and standardize continuous variable, 然后看boolean的frequency，是不是skewed. 还要看有没有multicollinearity issue,和heteroskedasticity, 所有的assumption满足了才能做regression， 他又问有collinearity怎么办，不是perfect collinearity, 我说用ridge regression..他就没有再问了，感觉答的不是太好吧。

---------- 3
http://www.mitbbs.com/article_t/Statistics/31263321.html
一面：
1. 他简单介绍了statistician@Google，主要有三个组，search quality和ads quality做data mining; NY的engineering组做machine learning; marketing组就是做marketing支持了。
2. 所学课程，照读一遍就OK。
3. 用的软件，不出所料，他们用R + Python.
4. 介绍一个你做过的project。我第一次在描绘project的时候被提很新颖的建议，还是从economics借的idea，真的很意外，不过想想interviewer是tenured professor就不以为怪了。
5. 我终于被问问题了！！！之前几个面试包括实习的电面都没有被问过，一道也没有！
a) 一道是关于多重共线性的。OMG，qualify之后就没看过，问了ridge regression对bias/variance的影响。相关方法。我答的不好，只答了lasso等shrinkage method。看他意思还有其他。
b) 实践问题。建个model，比较marketing promoting前后。感觉他们做的很有意思，我给点关键词大家自己去研究/准备吧。keywords: logistic regression, randomize trial, random forest, nonresponse, propensity score. 我答的太慢，他省略了若干题。感觉interviewer是个很好的老师，喜欢一题一题的深入问下去，跟着他一个一个情景的想，很有意思。只可惜本人水平差矣。

二面：
1. what are the modeling techniques you have used?
2. what's the difference btw fixed effect model and mixed model? fixed effect Vs. random effect, how would you choose in practice/in your projects.
3. what are the techniques you use in longitudinal models/ survival models.
4. lots of questions about GLM: error structure, link function, estimation method, etc.

5. experience about SAS and R
6. experience of data manipulation
7. say you built a linear model, how to detect 异方差? suppose you saw an increasing pattern on your residue plot, how can you modify your model.
8. classification techniques you experienced.
9. experience with times series

new problem set, did not finish, probably b/c my poor answers ;p To be honest, I wasn't quite catch the problem. And I personally do not think the problem is well state.

suppose we want to modify search results by IP locations and compare click thru rate.

1. divide into two groups urban or rural. then compare. please comments on the design.
> I did not quite get here, which is very bad since this is the very first question, the very basic building block. my understanding is urban was assigned to trtm, rural was assigned to control. so I said it was biased design.

2. what would you recommend.
> randomize. It was actually amazing that he said they can do that !!! I probably should ask the randomize scheme they use :-)

3. suppose you got the result about trtm group and control group, how will you conduct the test?
> I was really confused here, wasn't that just H0:p1 = p2 Vs H1 p1>p2, then the simple proportion test. I asked whether we know n1, n2. he did not answer and skipped to next question.

4. if you find the test power is too low, how would you explain "test power is too low" to product manager, what's your suggestion to improve power?
> I said increase sample size, or better matching trtm/control group, make two groups as similar as possible in other attributes. Does not seems to be the right answer, since he quickly finished the problem and said those problems are tricky…

---------- 4
http://www.mitbbs.com/article_t/Statistics/31277607.html
上来就问什么是logistic回归。解释了一遍。然后问怎么做参数估计.印象当中有3种方法，可是及不出来。只好说MLE，score method什么的。 然后问我说MLE又怎么做，我记得好像用拉格让日乘子什么的，拉格朗日还不知道英文怎么说。还有New-Rapson什么的。记不清了 。

后面一个问题是，如果logistic回归自变量x不是线性的，怎么办？这个问题我也不知道。只好说那就用多项式回归，然后问我说多项式回归有什么risk？还有多项式回归怎么选order

下一个问题好像是contingency table的问题：
先问我怎么来design看人们喜欢google 地图还是别的地图。我说那就做个survey。然后她问我怎么处理这个survey数据。比如n个人做了这个实验。这样看上去是contingency table啊，可是刚才面试的时候我也慌了。我说的是用开方检验。

接着问的是，除了记录他们喜欢那个地图以外，还有什么要记录的？我说的是还有response time，客户体验，想说的挺多的就是不知道怎么用英语表达出来。

回答的太撮了，英语也说的磕磕绊绊的。然后她问我有没有问题，我问了两问题。结果她又说还有问题。

下一个问题，这个问题我也没怎么弄明白：
使用chrome，假如有记录从开始到现在people 使用chrome上的搜索引擎的记录，那么怎么分析这个data，客户是不是更喜欢google？

刚开始我觉得这个问题跟google地图那个一样。她说不一样，这个是有各个时间的记录。我说这不就是时间序列么？然后问了我一些时间序列的东西。

还是这个问题，能不能做回归？y是记录值，x是时间。 我说我觉得不可以。大牛来说说这种数据该怎么分析？

回答的挺不好的。感觉除了统计的一些知识外，还要考虑怎么设计实验。基本上给的问题都是我们想检验某个结果，你怎么去设计实验，怎么收集data，怎么分析？

---------- 5
http://www.mitbbs.com/article_t/JobHunting/32671827.html
Multicolinearity问题:
首先，最简单的模型 Y=Xb+e 的LS解是b_hat=(X'X)^(-1)X'Y， var (b_hat)=(X'X)^(-1) sigma^2.

问题：什么是Multicolinearity,
答：如果承认X是rv，才能用"correlated "，否则只能用比较数学的linear dependent，not of full column rank这种术语。

Multicolinearity又分2种，multicolinearity 和perfect multicolinearity，分别对应的是X的column vectors 是 nearly linear dependent和 linear dependent(not full ranked)，分别对应的结果就是 (X'X) 是ill-conditioned 和singular.，前者是(X'X)^(-1)存在但是norm 非常大，后者是(X'X)^(-1)根本就不存在必须用generalized/pseudo inverse 来解决，记作(X'X)^(-)。

问题:存在multicolinearity的时候 b_hat 是unbiased的吗？

答案：如果不是perfect multicolinearity，(X'X)^(-1)总是存在的。
E(b_hat)=E((X'X)^(-1)X'Y)=(X'X)^(-1)X'Xb=b，所以unbiased.

如果是perfect multicolinearity，(X'X)^(-1)不存在只能用(X'X)^(-),这个有无穷多，所以E(b_hat)= (X'X)^(-)(X'X)b根据不同的(X'X)^(-)的选法也是无穷多的，根本就不是unbiased。

但是E(y_hat)=E(Xb_hat)=X(X'X)^(-)(X'X)b=XX^(+)Xb=Xb=E(Y)，跟(X'X)^(-)的选法无关(Moore-Penrose theorem)， 所以y_hat=Xb_hat对Xb永远是unbiased的。我说这个的时候面试官发笑我很无语啊。。。

问题： multicolinearity的treatment?

很多。最著名的是ridge regression，但是ridge regression只是 Tikhonov Regularization的special case, 我说出来Tikhonov regularization 大家一副我在扯淡的表情的我实在很郁闷。。。