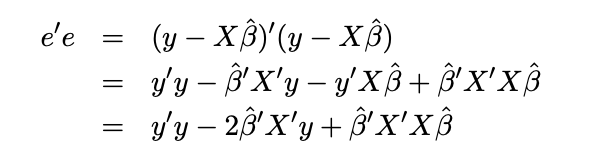
****

**square of a vector is defined as square of vector is defined as dot product of a vector with itself, 所以一定要注意dimension是1x1**

**以及每一项的先后顺序，特别是B和X的顺序**

**Hypothesis Testing 注意one-tail 还是two-tail**

**Decision Tree 薄弱环节**

Cynthia Rudin Notes, ESL

**区分generalized least squares 和generalized linear models**

**linear regression, 从推系数，到regularized lr，再到gener‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍alized least squares，再到generalized linear models全问了一遍**

<https://www.1point3acres.com/bbs/thread-613195-1-1.html>

**bias variance tradeoff analysis**

<https://www.1point3acres.com/bbs/thread-830501-1-1.html>

**Linear Regression 圣经**

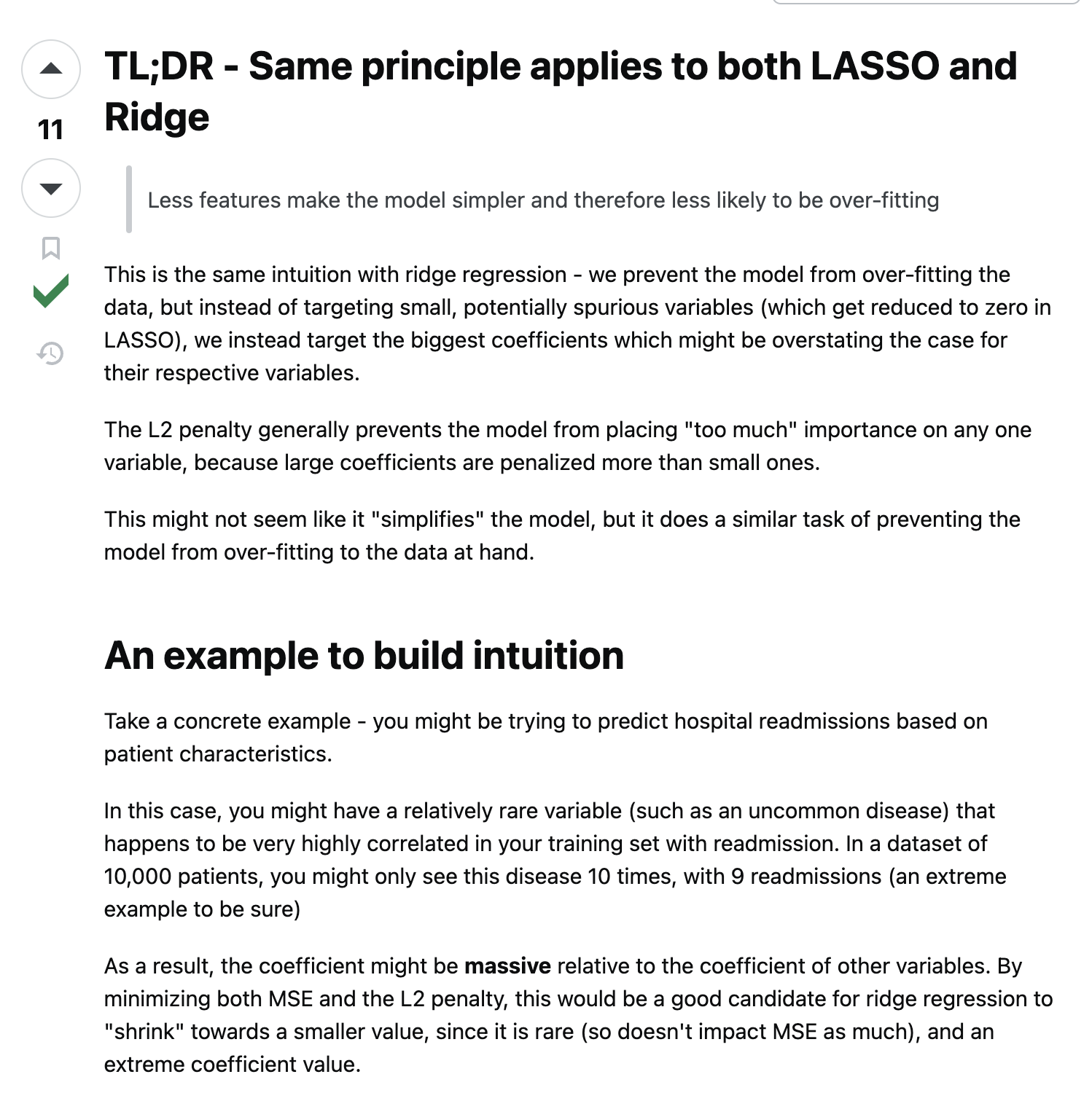
**这篇好像是median regression最早的出处**

**- How to solve multiple linear regression (我用normal equation解释的 然后给了OLS的close form)**

**- What if model works very well on training but poorly on testing (CV)**

**- What if with CV it still works poorly on testing? (regularization)**

**- How are L1, L2 reg different? How would resulting coefficients look? Why is that? Why would we want coefficient with small absolute value?**

[https://stats.stackexchange.com/questions/267772/in-ridge-regression-and-lasso-why-smaller-beta-would-be-better](https://stats.stackexchange.com/questions/267772/in-ridge-regression-and-lasso-why-smaller-beta-would-be-better)

**- What if model just performs very poorly? (I answered maybe the data just doesn’t follow linear regression assumptions and explained the assumptions)**

**- What if the noises are independent and have the same distribution but are just very very big? (I answered median regression)**

**- Why is median regression better in this case?**

**- If median regression is so good, why would we still use OLS? (这里我不太确定要答什么 随便讲了讲 最后他告诉我median regression converges slowly during gradient decent)**

**最后linear regression终于问完了**

**- If data doesn’t really have linearity, what do we do? (I answered we can use non-linear models like random forest or DL)**

**- He said “pick one” (I picked random forest and explained how it works)**

<https://www.1point3acres.com/bbs/thread-754686-1-1.html>

**Linear Regression (Lasso, Ridge, regularization), assumptions, why using mean squared error? (based on normality assumption, because we want to maximize the maximum likelihood),**

<https://www.1point3acres.com/bbs/thread-818794-1-1.html>

**transfer lasso to quadratic programming problem**

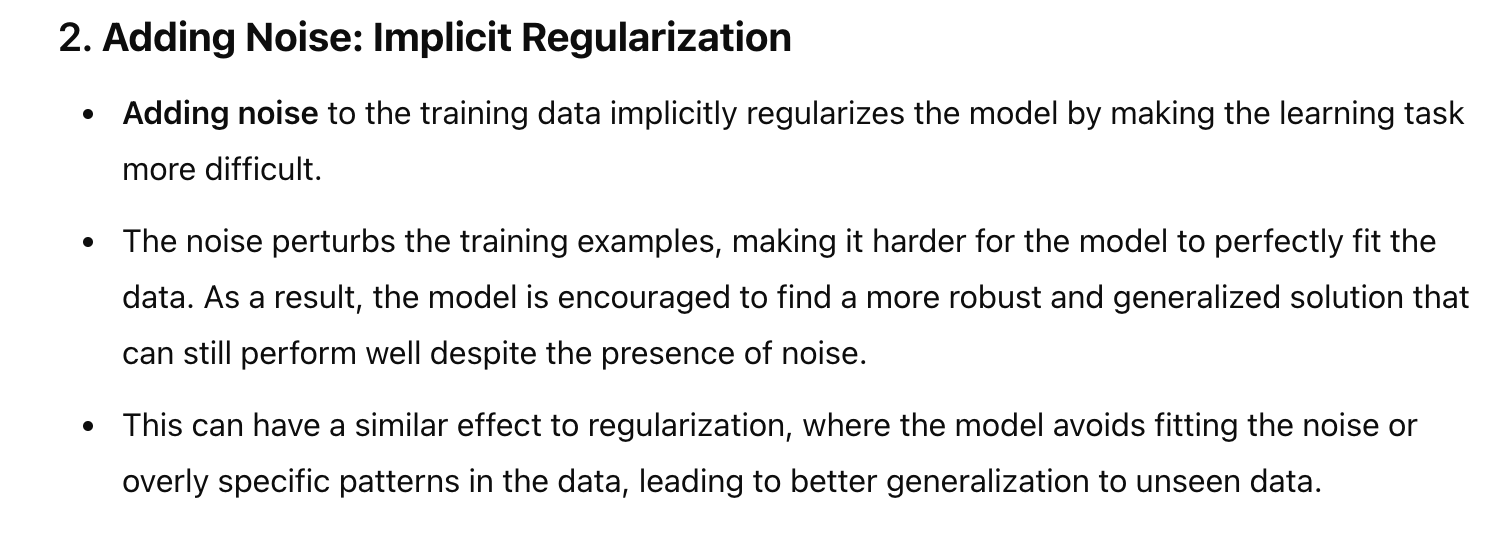
<https://www.1point3acres.com/bbs/thread-818794-1-1.html>

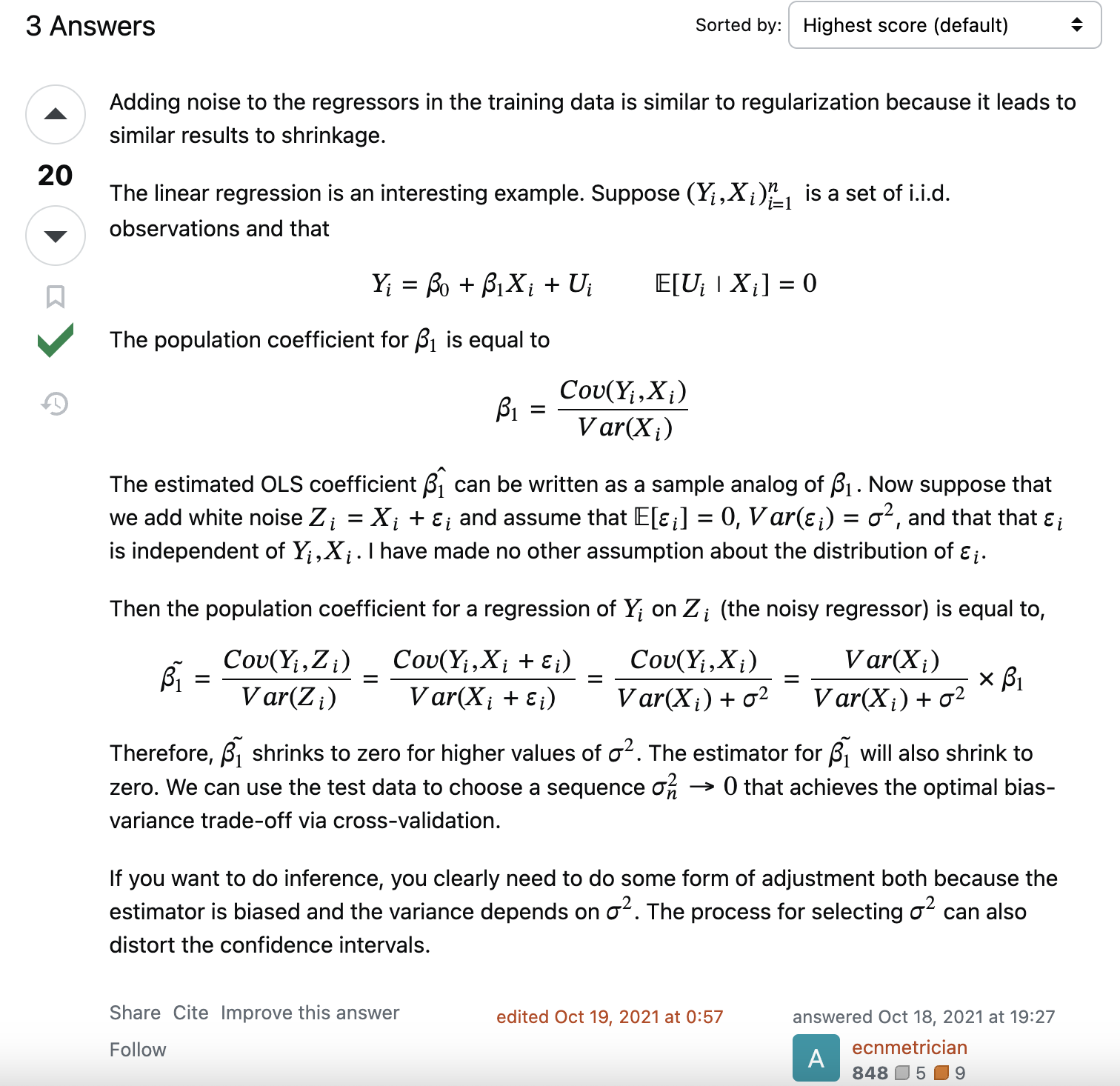
EE364b [L1 methods for convex-cardinality problems](https://web.stanford.edu/class/ee364b/lectures/l1_slides.pdf) notes, search for LASSO

**Linear Rergression 如果两个变量0.999相关，权重会如何分配，如果是lasso 和 ridge分别会如何，实践中如何选择保留哪个变量。如果在x加上噪声，会怎么样，分别从直观感受和推导上说，最后一直说到给x加噪声可以起到和正则化同样的作用**

<https://www.1point3acres.com/bbs/thread-710173-1-1.html>

<https://stats.stackexchange.com/questions/548756/how-is-adding-noise-to-training-data-equivalent-to-regularization>





**linear regression如果double数据点，对coef(也叫beta)，tstats和R2会怎么变**

<https://www.1point3acres.com/bbs/thread-775511-1-1.html>

<https://www.1point3acres.com/bbs/thread-925252-1-1.html>

**regression如果duplicate了所有的数据， estimate (也叫mean) 会怎么变? 以及variance会怎么变?**

<https://www.1point3acres.com/bbs/thread-686326-1-1.html>

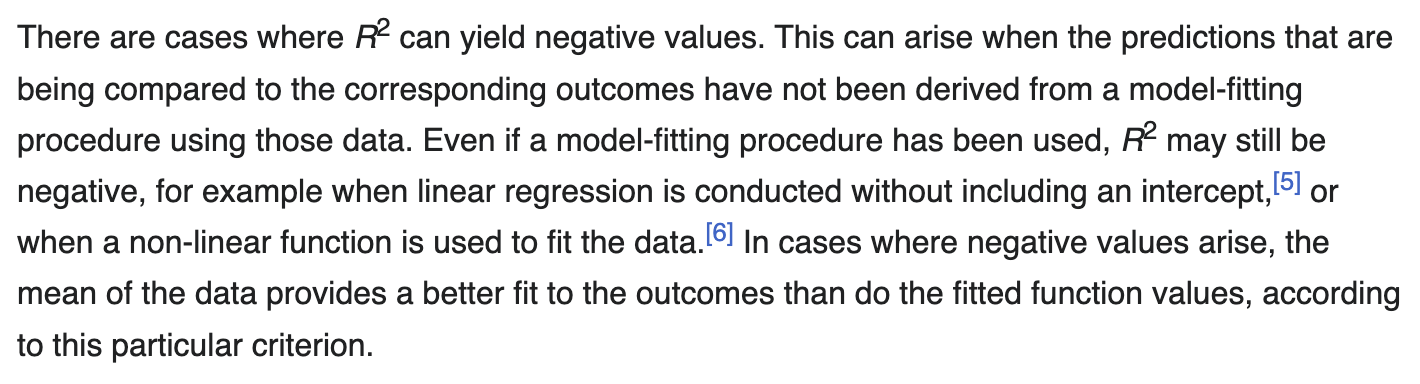
<https://www.1point3acres.com/bbs/thread-710173-1-1.html>

**Linear Regression 的R^2何时为负数**

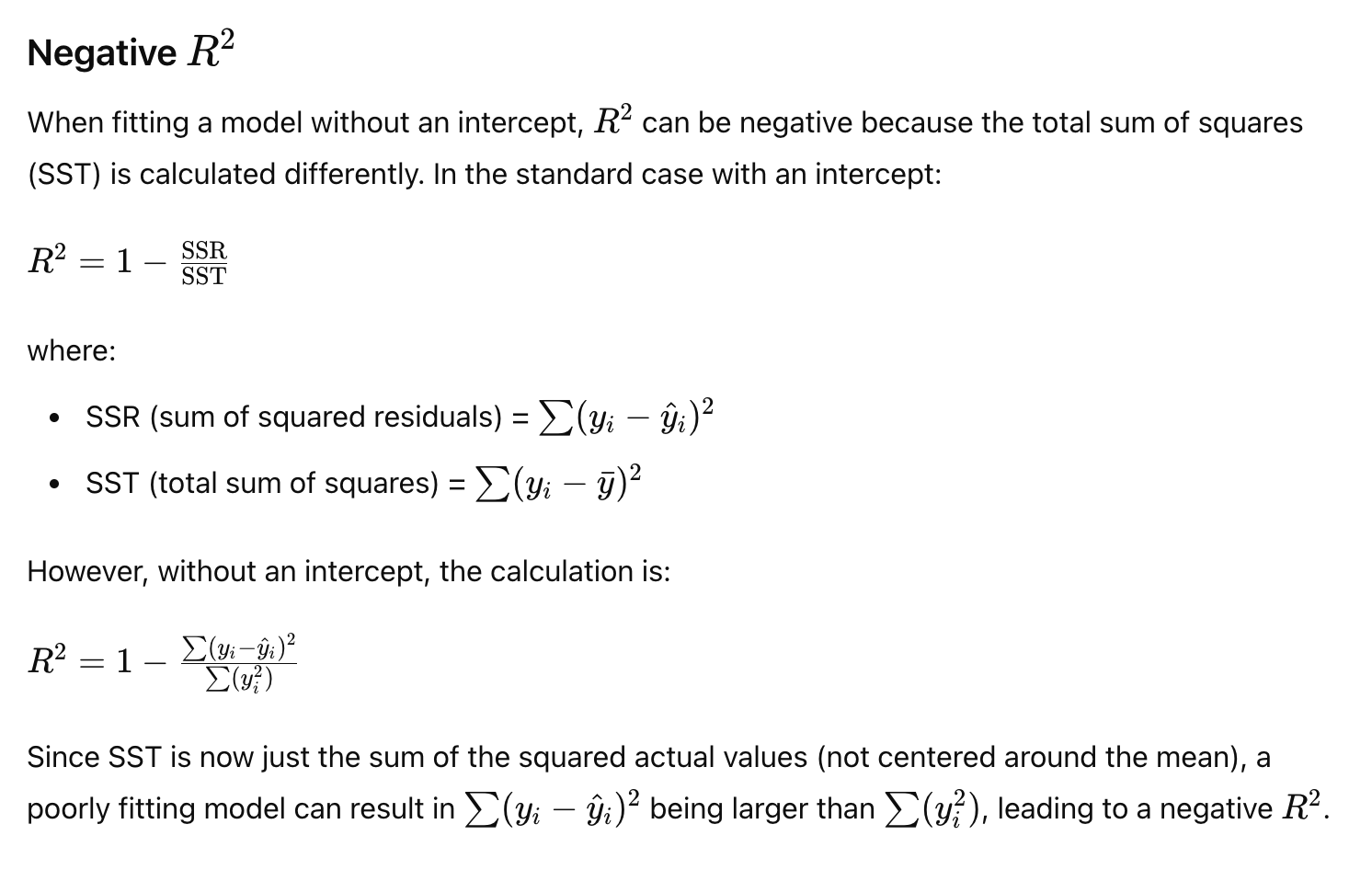
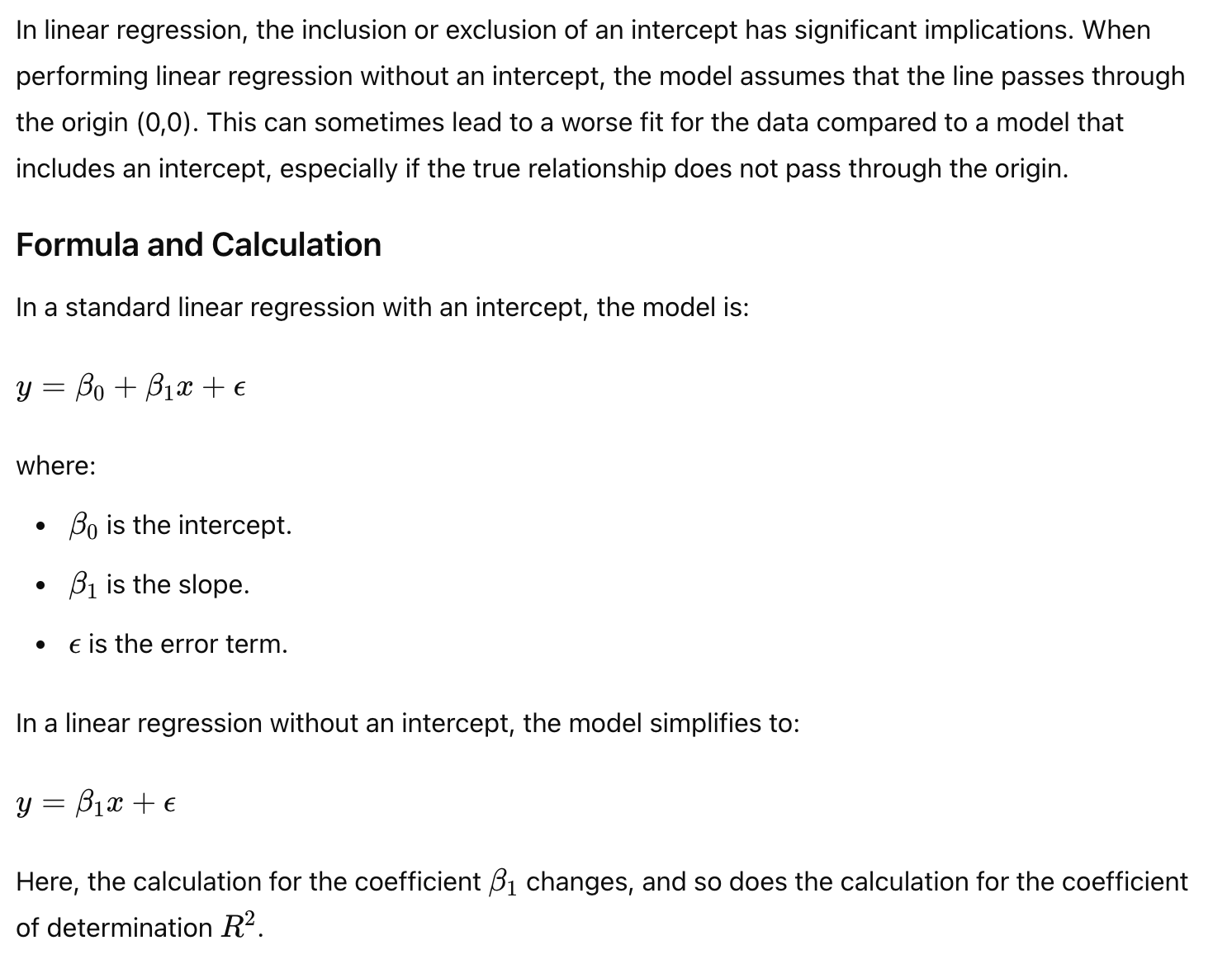
<https://www.1point3acres.com/bbs/thread-1025314-1-1.html>

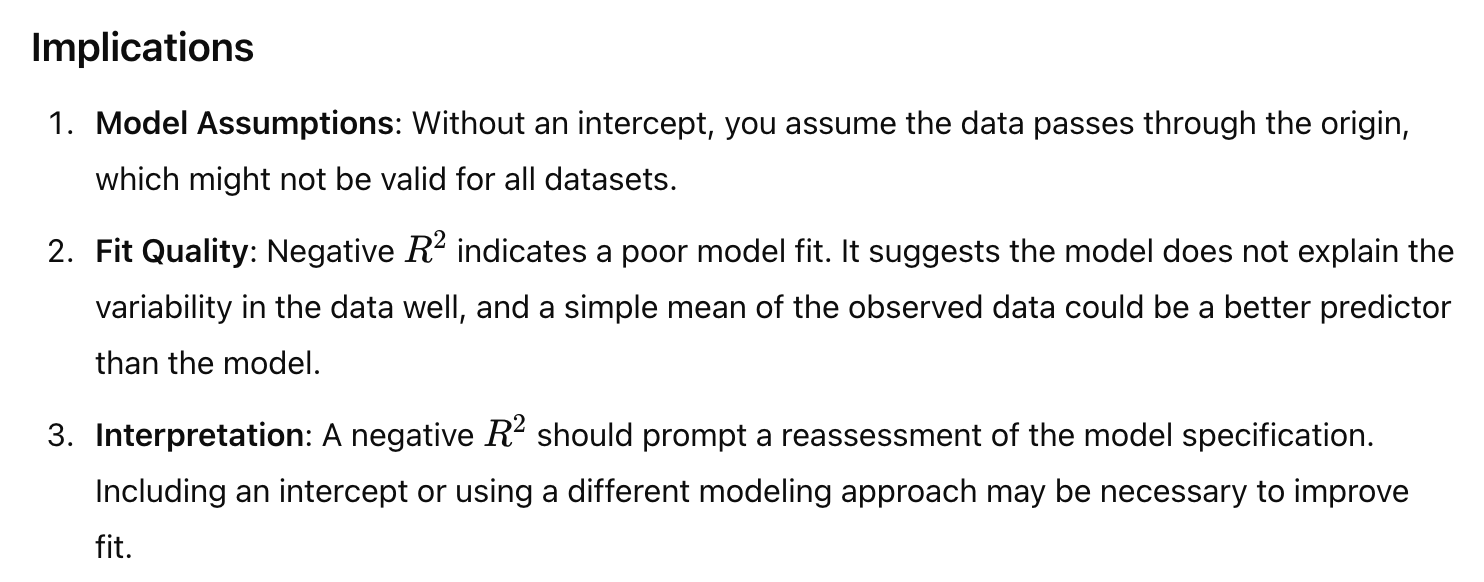
下面这段来自wiki <https://en.wikipedia.org/wiki/Coefficient_of_determination>

**比较重要的是R^2为负数时，mean of data providers a better fit to the outcomes than the fitted function values**



之前确实没太想过linear regression without an intercept的情形





**当数据的特征改变的时候ridge regression的lambda 会怎么变**

<https://www.1point3acres.com/bbs/thread-1018348-1-1.html>

**重点是more complicated model, needs a higher lambda**

**highly correlated features need a higher lambda**

**if important features moved, model is now under-regularized, and thus needs a higher lambda**

When the features of a dataset change, the effect on the Ridge regression's lambda (regularization parameter) depends on the nature of the change in the features. Here’s how it typically works:

1. \*\*Feature Scaling or Normalization\*\*:

- Ridge regression is sensitive to the scale of the features because it involves the sum of squared coefficients. If the features are not normalized, the lambda might need to be adjusted.

- \*\*If features are scaled up or down\*\*, the effect of the regularization term changes. For instance, if features are scaled up, the regularization might become less effective, and you might need a higher lambda to maintain the same level of regularization. Conversely, scaling features down might require a lower lambda.

- \*\*Normalization\*\* (standardizing features to have a mean of zero and a standard deviation of one) is common practice in Ridge regression to ensure that lambda has a consistent effect across features.

2. \*\*Addition of Features\*\*:

- If new features are added to the model, the complexity of the model increases. This might require an adjustment of lambda to maintain a balance between model complexity and regularization.

- \*\*Adding highly correlated features\*\* might make the model more susceptible to multicollinearity. Ridge regression is often used to mitigate multicollinearity, so adding such features could lead to an increase in the optimal lambda.

3. \*\*Removal of Features\*\*:

- Removing features can reduce model complexity, possibly allowing for a lower lambda. If important features are removed, the model might become under-regularized, requiring an increase in lambda to control the variance.

- If redundant or highly correlated features are removed, it might allow for a decrease in lambda since the model would be less prone to overfitting.

4. \*\*Change in Feature Relationships\*\*:

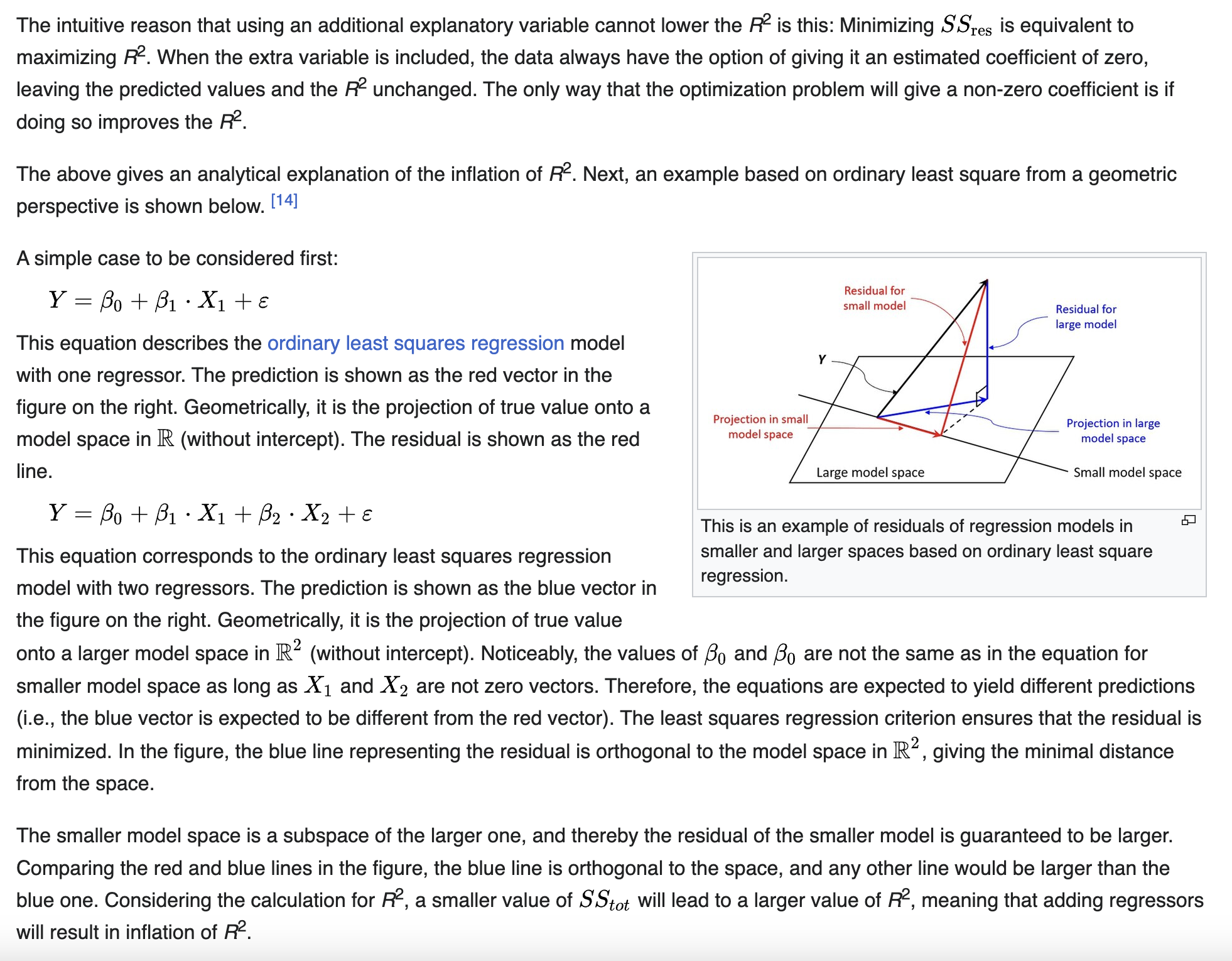
- If the relationships between features change (e.g., interactions between features are introduced or removed), the regularization effect might need to be adjusted by changing lambda.

### Summary:

The lambda in Ridge regression might need to be adjusted based on how the features are scaled, added, removed, or modified. Proper feature scaling is essential to ensure that lambda has a consistent impact, and changes in feature relationships may require re-evaluation of the optimal lambda.

**R^2和features的关系，数目越多R^2越大吗？如果假定feature没有严重multicollearity，对于一个n个data point的模型，估计最多多少合理？**

Inflation of R^2的analytical和geometric interpretation都挺有趣的, **R^2的wiki界面全文背诵**；第二问就lasso和ridge呗



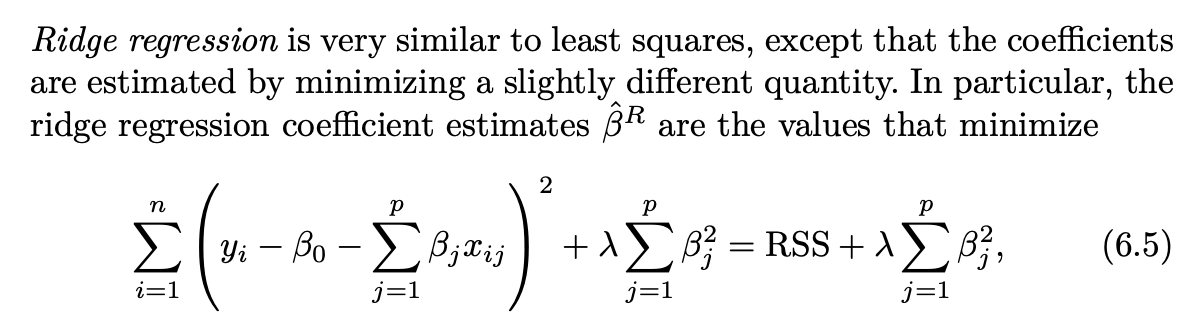
<https://www.1point3acres.com/bbs/thread-953725-1-1.html>

**Ridge 对 R2 有什么影响**

<https://www.1point3acres.com/bbs/thread-929982-1-1.html>

**ESL P64, ridge regression makes X^TX nonsingular by adding a constant even X is not of full rank**

ridge regression typically shows a **trade-off between bias and variance**. It might have a olwer R^2 on the training data but can perform better on unseen test data compared to OLS, leading to a higher R^2 on the test data.



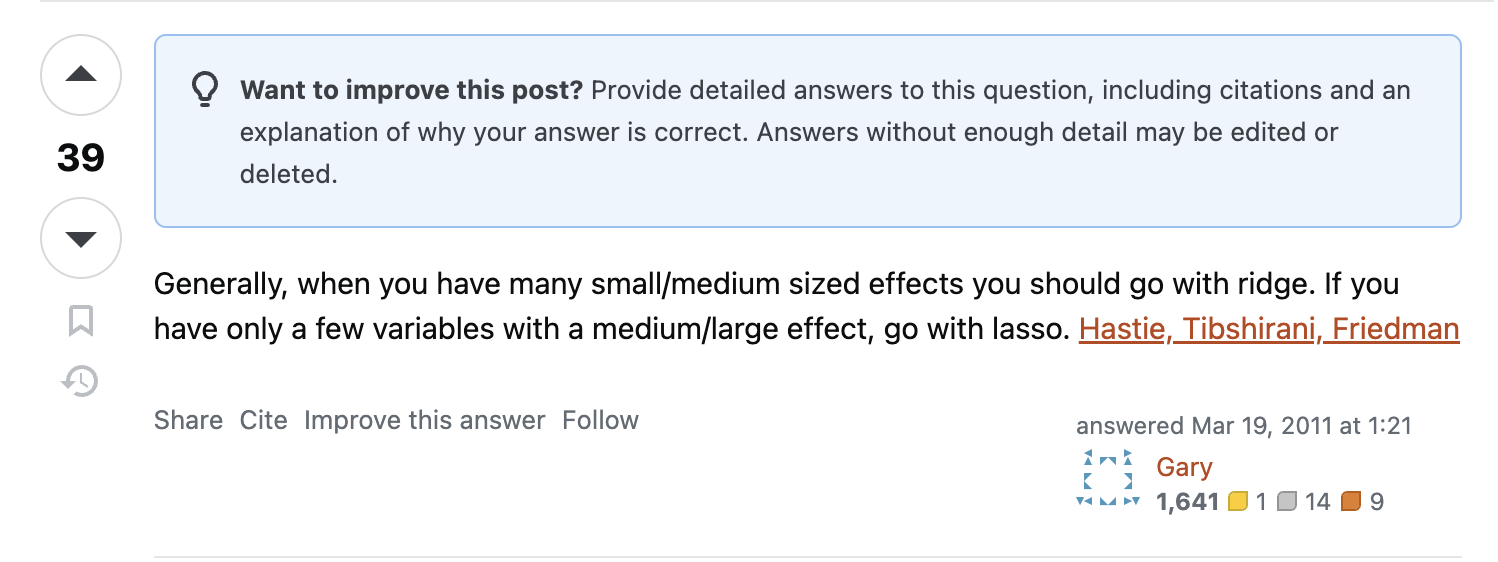
Ridge Regression都minimize这玩意了，那肯定不minimize RSS，自然training set的R^2会更大

**linear regression. How the coefficient will change if we duplicate a feature? How the coefficient will change if we apply a regularization to the ‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍duplicated dataset?**

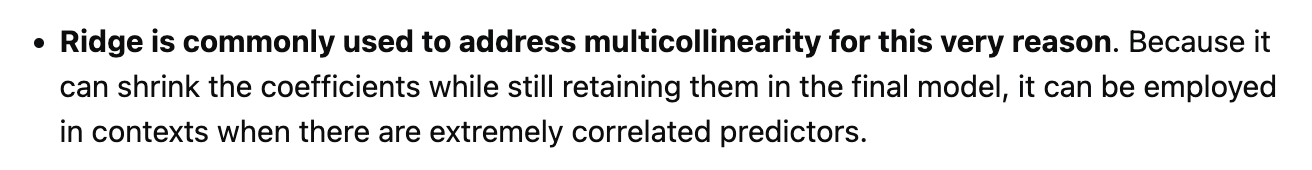
<https://www.1point3acres.com/bbs/thread-851024-1-1.html>

**什么时候用Ridge，什么时候用Lasso**

<https://www.1point3acres.com/bbs/thread-892035-1-1.html>



**这个回答抓住了重点**

****

还有就是ridge能更好handle multicollinearity

**另外，还可以用elastic net**

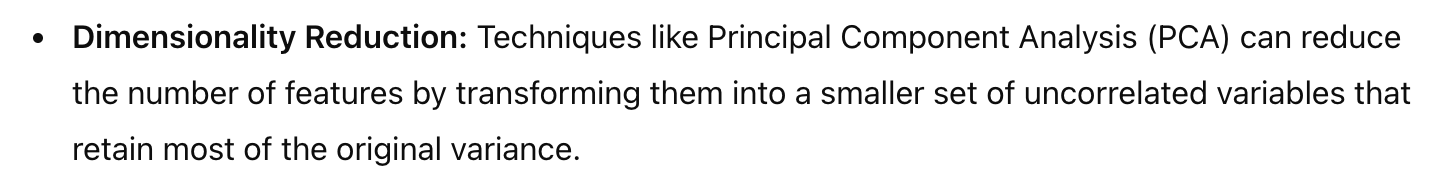
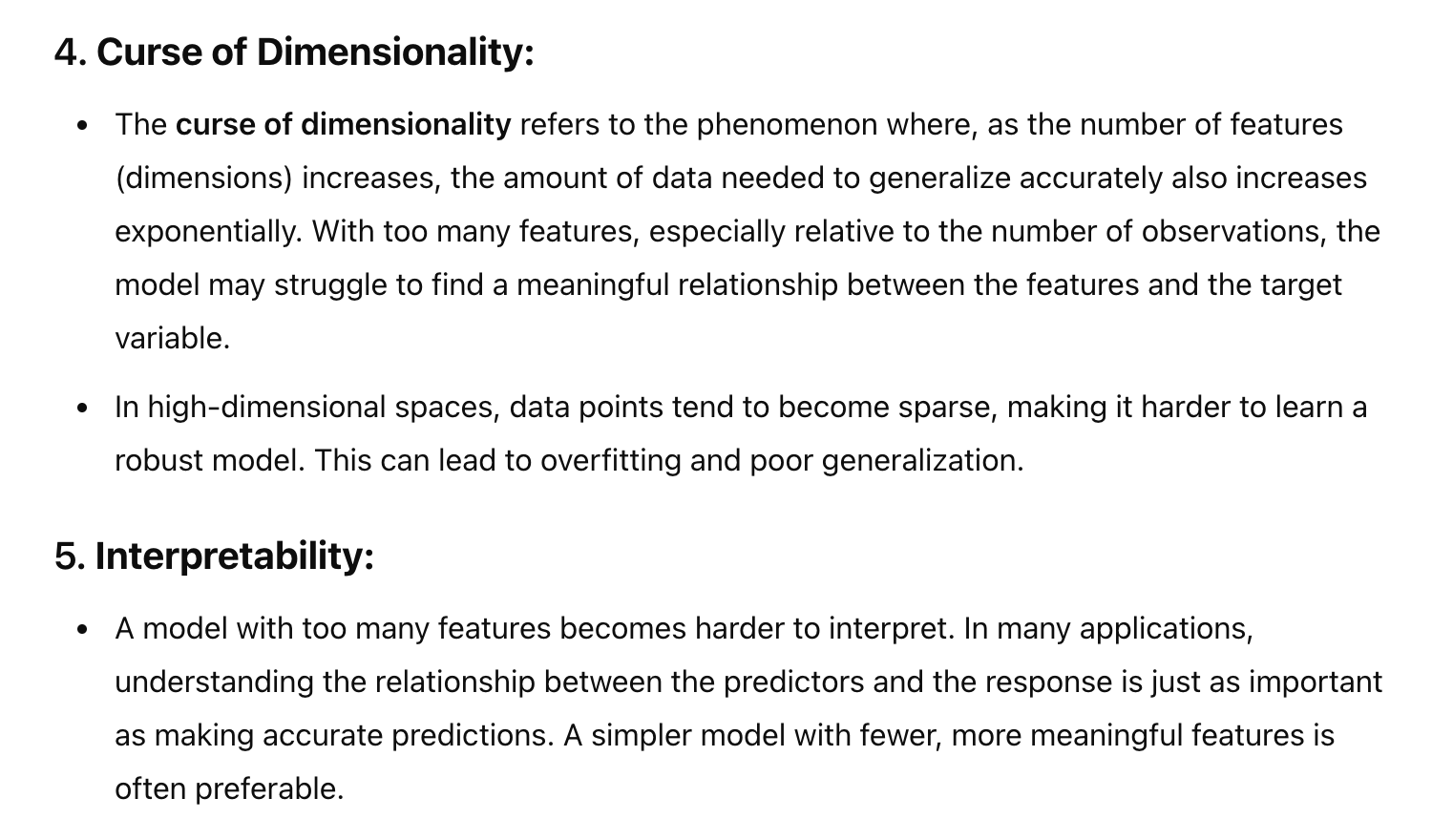
**regression feature数量太多了会发生什么问题?**

直接问chatgpt: “what happens if there are too many features in linear regression”

1. overfitting
2. multicollinearity
3. Increased Model Complexity
4. **Curse of Dimensionality**
5. **Interepretability**
6. Increased risk of irrelevant features

Solutions:

* Feature Selection
* PCA
* Cross Validation



<https://www.1point3acres.com/bbs/thread-686326-1-1.html>

**Precision Recall**

Cynthia Rudin Notes

**OLS close form矩阵太大怎么办，存不下来，(X'X)这一项我怎么计算：**

<https://www.1point3acres.com/bbs/thread-870889-1-1.html>

我覺得1. 想問的是 kernel method. 假設X: NxM. N rows, M columns (features). 假設你的data 很high dimension, aka M 超大, 那x^Tx 就是MxM. 要invert超麻煩. 如果使用kernel trick, aka Gram matrix K = XX^T. dimension 就是NxN. 當 N << M, 用kernel trick 會比較快 因為要invert的matrix dimension比較小.

我给了两个方法 1： PCA 2：SGD

说不是，他要的准确方法，我说那你就算一下每一个的cov然后取一下避免重复计算，OLS的矩阵问题一直是比较难搞的东西，我甚至还想说难道用LU？还说不是，我笑了，我说你说说，他说你写下来看看，sum来sum去的不知道说什么。

**相似问题：suppose the matrix is large as we have many features, how to speed up the calculation?**

<https://www.1point3acres.com/bbs/thread-664696-1-1.html>

**Solve least square solution in an online manner**

Recursive Least Squares Filter

这算法有点太advanced了，又复杂记不住又不intuitive,搞不清大体motivation和design

原题似乎也没有问很深入

<https://en.wikipedia.org/wiki/Recursive_least_squares_filter>

<https://www.1point3acres.com/bbs/thread-1020326-1-1.html>

**linear regression，问了l1 和 l2 norm的理解，**

<https://www.1point3acres.com/bbs/thread-695772-1-1.html>

# **Jarque–Bera test**

# **Cook's distance**

**p value, bootstrap**

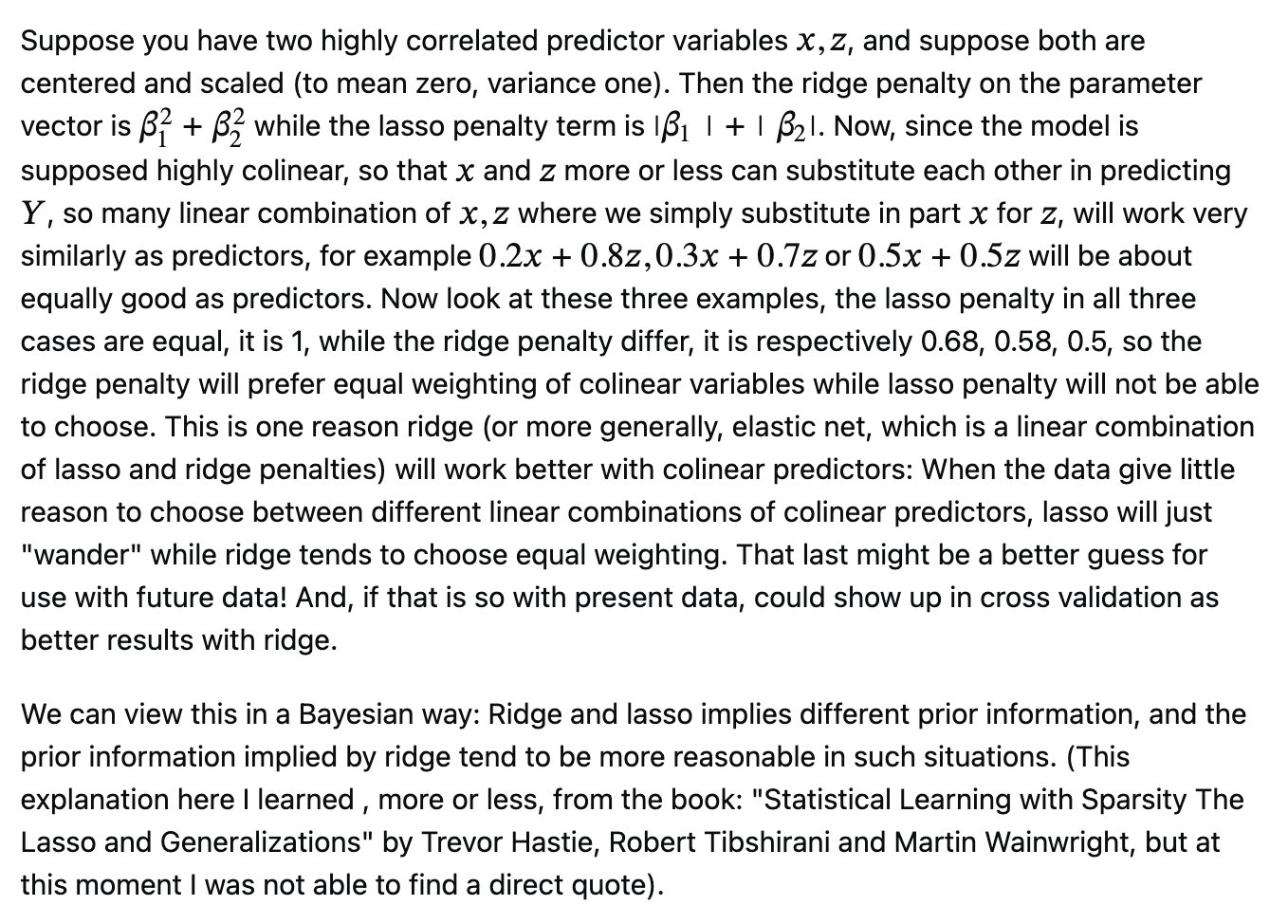
**p value, biased due to outlier**

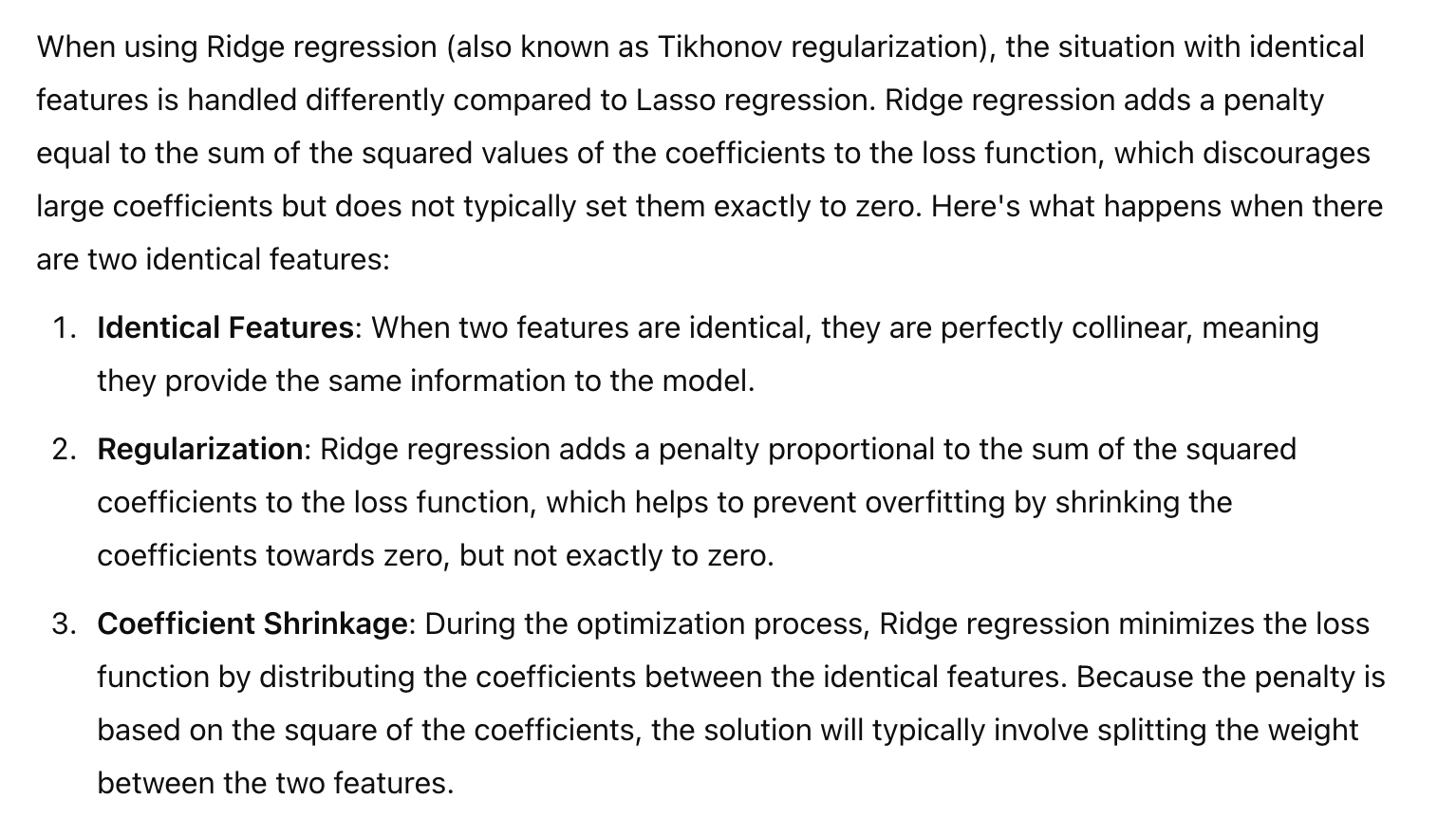
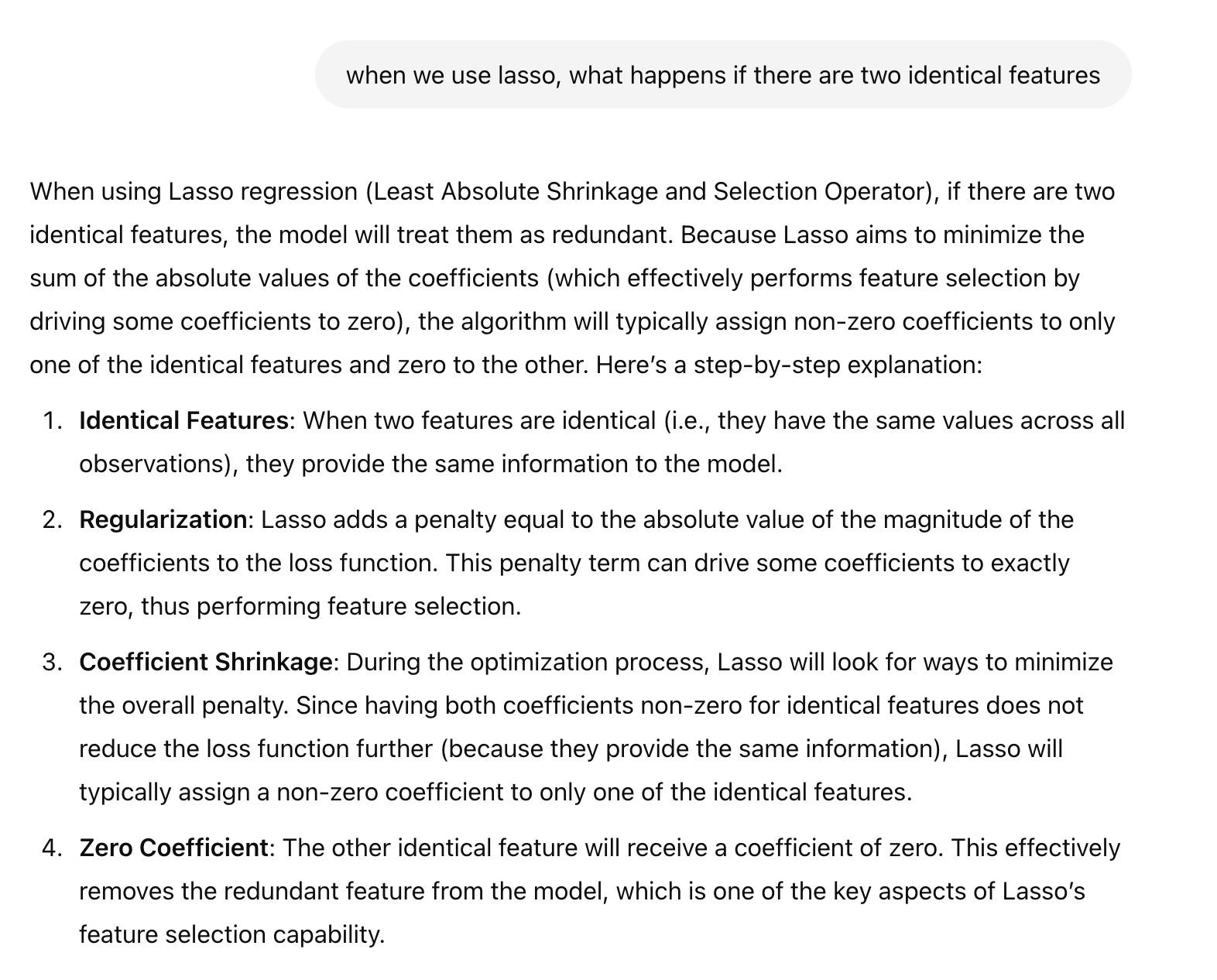
**先问了一个linear regression的问题，假设有两个variable duplicate，问会怎么样。接着又问了如果（用这个duplicated variable dataset) 做ridge regression，weights应该是什么样**

至于lasso会keep哪个，应该是很随机的，取决于具体implementation; 而且可以反问，既然都duplicate了，那keep哪个drop哪个还有所谓嘛

<https://www.1point3acres.com/bbs/thread-674785-1-1.html>

<https://stats.stackexchange.com/questions/264016/why-lasso-or-elasticnet-perform-better-than-ridge-when-the-features-are-correlat/264118#264118>





**Given an NxP matrix representing data collected, how to reduce dimensionality, aka simplify data (analytically and mathematically).**

答案是关于rank，image，orthogonality/gram sc‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍hmidt之类的分析。

楼主赞,请问第三题如果分析rank 和orthogonality? 是用PCA吗

没学过PCA，但是查了一下，跟面试官想要的答案很相似，就是找到一些vector，之间的correlation降到最低。

<https://www.1point3acres.com/bbs/thread-680077-1-1.html>

**Linear Regression, 多加feature对R值得影响**

原始笔记应该有

Inflation of R^2, as well as increasing test error (STAT521 Lec 7)

**时间序列相关，ARIMA全家桶？AR(1) stationary的条件是什么？**<https://www.1point3acres.com/bbs/thread-680805-1-1.html>

**【悬案】又一个经典题目是 regression y～x x～y 两个系数乘机小于1. follow up: 如果是 median regression 就是loss 是absolute v‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍alue， 两个系数有啥关系**

<https://www.1point3acres.com/bbs/thread-714998-1-1.html>

有个回答是：

用median代替mean，co-median代替covariance，答案还是-1到1之间。见paper：<https://www.ism.ac.jp/editsec/aism/pdf/049_4_0615.pdf>

**蒙特卡洛怎么work， 如何减少variance**

虽然是个个人向问题，但了解下也是好的

<https://www.1point3acres.com/bbs/thread-719697-1-1.html>

**问简历和Time series 预测有关的项目，如何detrend, 如何选择moving averag‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍e feature**

虽然是个个人向问题，但了解下也是好的

<https://www.1point3acres.com/bbs/thread-719697-1-1.html>

**先问了一些estimator的性质，**

<https://www.1point3acres.com/bbs/thread-934609-1-1.html>

然后问weighted average as an unbiased estimator, what weights can make the estimator with least varianc‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍e?

<https://www.1point3acres.com/bbs/thread-934609-1-1.html>

实际上是一个Lagrangian multiplier，但是做的时候没想到，有点可惜，最后没完全证出来

先自己推下式子，搞清楚要解的优化问题是什么

**题目是如果假设有个data matrix X [dimension (n, k)]，k是features，Y是labels [dim (n, 1)]。现在想做ordinary linear regression，但是n非常非常大（不过k不大），然后现在有m个machines可以分担workload，要求我写一个mapreduce的算法算出相应的解。**

原题提示：结果小哥说想看看我怎么算closed form solution

(X^T X) ^{-1} X^T就是提示

先map reduce求出 k by k 的X^T X (只需要column i 的transpose 乘 column i 就行），然后随意求inverse，因为k很小

再map reduce乘X^T， 只不过这次用row j就是

具体的map reduce回头再说

<https://www.1point3acres.com/bbs/thread-676544-1-1.html>

**linear regression 解析解公式推导。 如果有一个大矩阵乘以一个小矩阵，无法fit into memory，如何更快？**

感觉和上面是同一道题

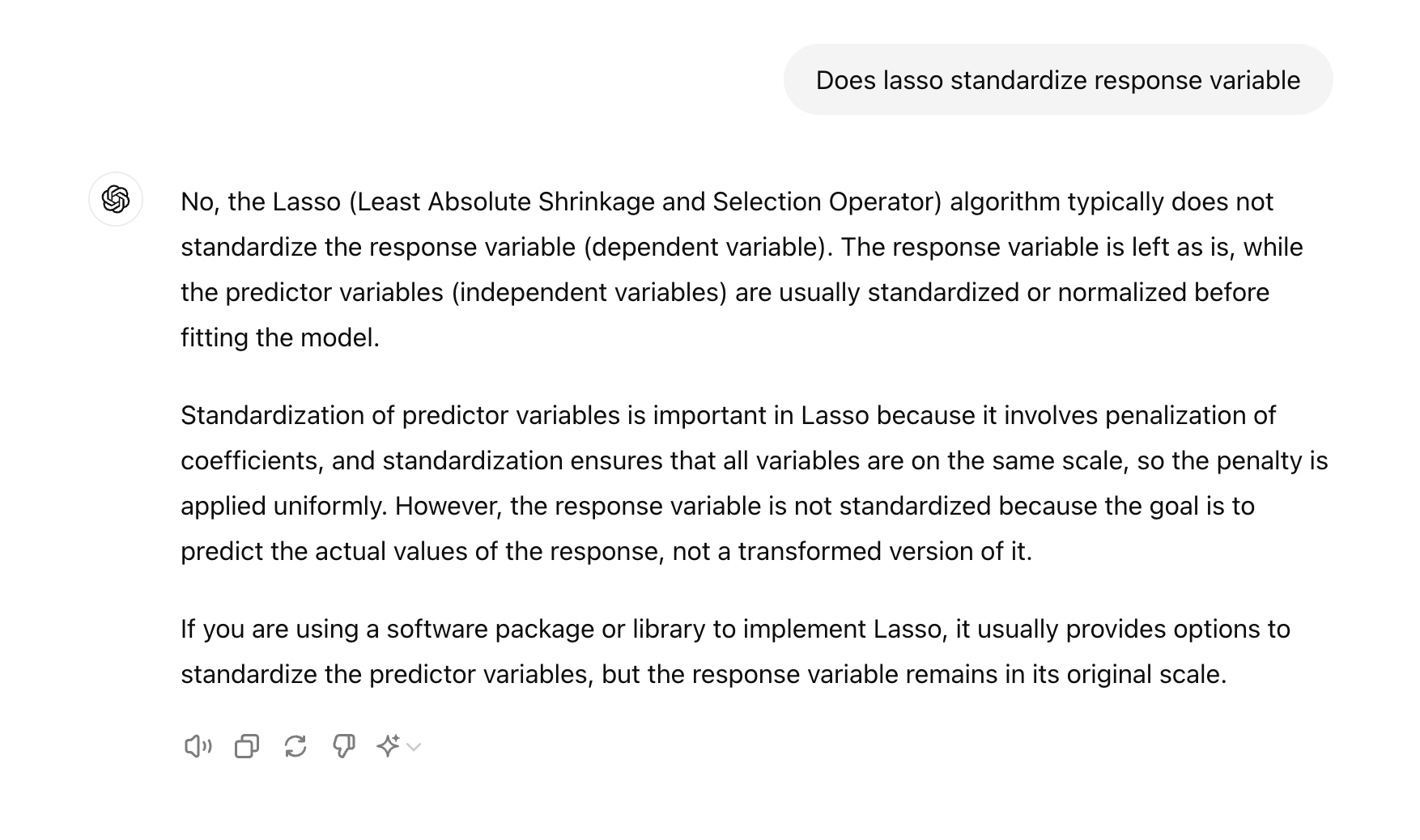
拆分成小矩阵乘，推公式

<https://www.1point3acres.com/bbs/thread-633629-1-1.html>

**然后问我lasso把一个predicotr variable的data都乘2，这个predictor的系数是会怎么变。好像这题还有一个变种是ridge regression会怎么变。**

旧note有详细解释, standardization, 不变

<https://www.1point3acres.com/bbs/thread-525249-1-1.html>



**有两个random variable，都遵从normal distribution, 用一个fit另一个做Linear regression, 这样得到两个beta, 他们的乘积有没有bound？（要求给出至少两种解释。有概率的, 还有几何的）**

independent也是Make sense的。题目故意很vague，需要自己去解读。你可以想象成两个vector，vector是由不断从一个distribution中采样产生的

总体没什么思路

<https://www.1point3acres.com/bbs/thread-633629-1-1.html>

**\item 如果\(X \thicksim N(0, \sigma\_{1}^{2}), \: Y \thicksim N(0, \sigma\_{2}^{2})\)。假设使用无intercept的regression，用\(Y\) 对\(X\) 做regression 获得系数\(\beta\_1\), 用\(X\) 对\(Y\) 做regression获得\(\beta\_2\)，问\(\beta\_1 \beta\_2\)的范围。**

<https://www.1point3acres.com/bbs/thread-816723-1-1.html>

**考虑Y～X，X和Y iid normal （n个sample）, 做linear regression，他们的R2 期望是多少。拓展是Y ~ X1 + X2 + ... +Xk, 仍然是iid，R2期望是多少。\**

列一下式子就行了，大概是k/n

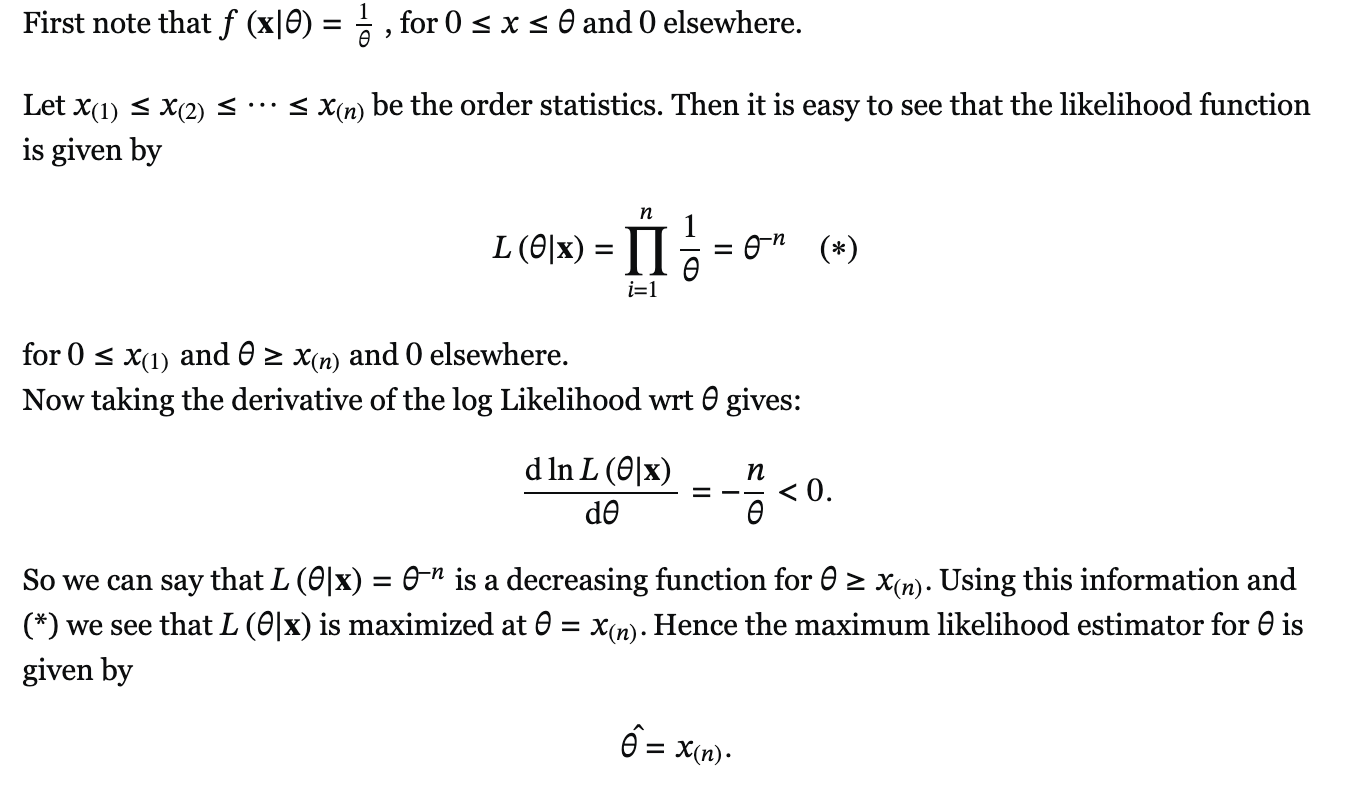
<https://www.1point3acres.com/bbs/thread-1026270-1-1.html>

**0到K的均匀分布，手推K的MLE，是否无偏**

这结果肯定biased

<https://www.1point3acres.com/bbs/thread-647068-1-1.html>

<https://math.stackexchange.com/questions/49543/maximum-estimator-method-more-known-as-mle-of-a-uniform-distribution>



**Open question: bayesian modeling of the probability of some rare events. What's your prior? What's the parametrized emission prob? How to derive the posterior**

<https://www.1point3acres.com/bbs/thread-925252-1-1.html>

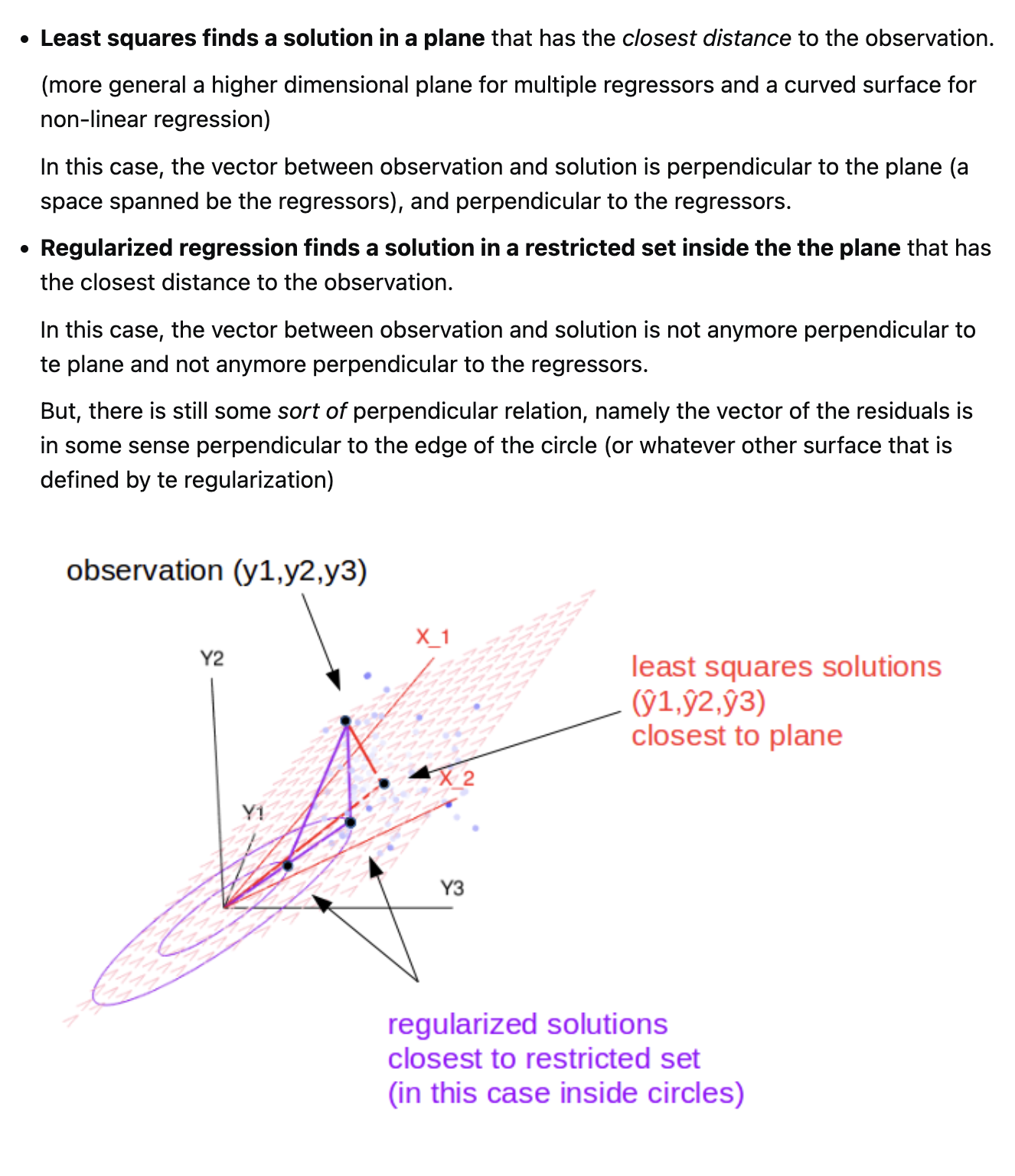
**Ridge regression: what preprocessing we need to do? Are the residuals orthgonal to predicted values?**

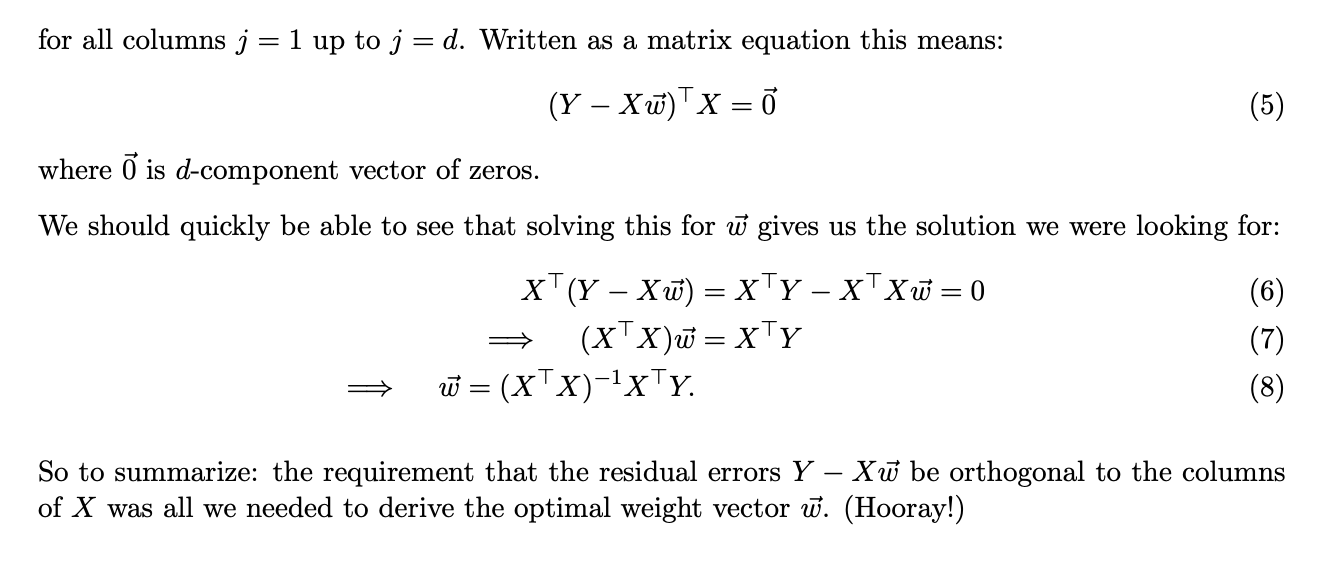
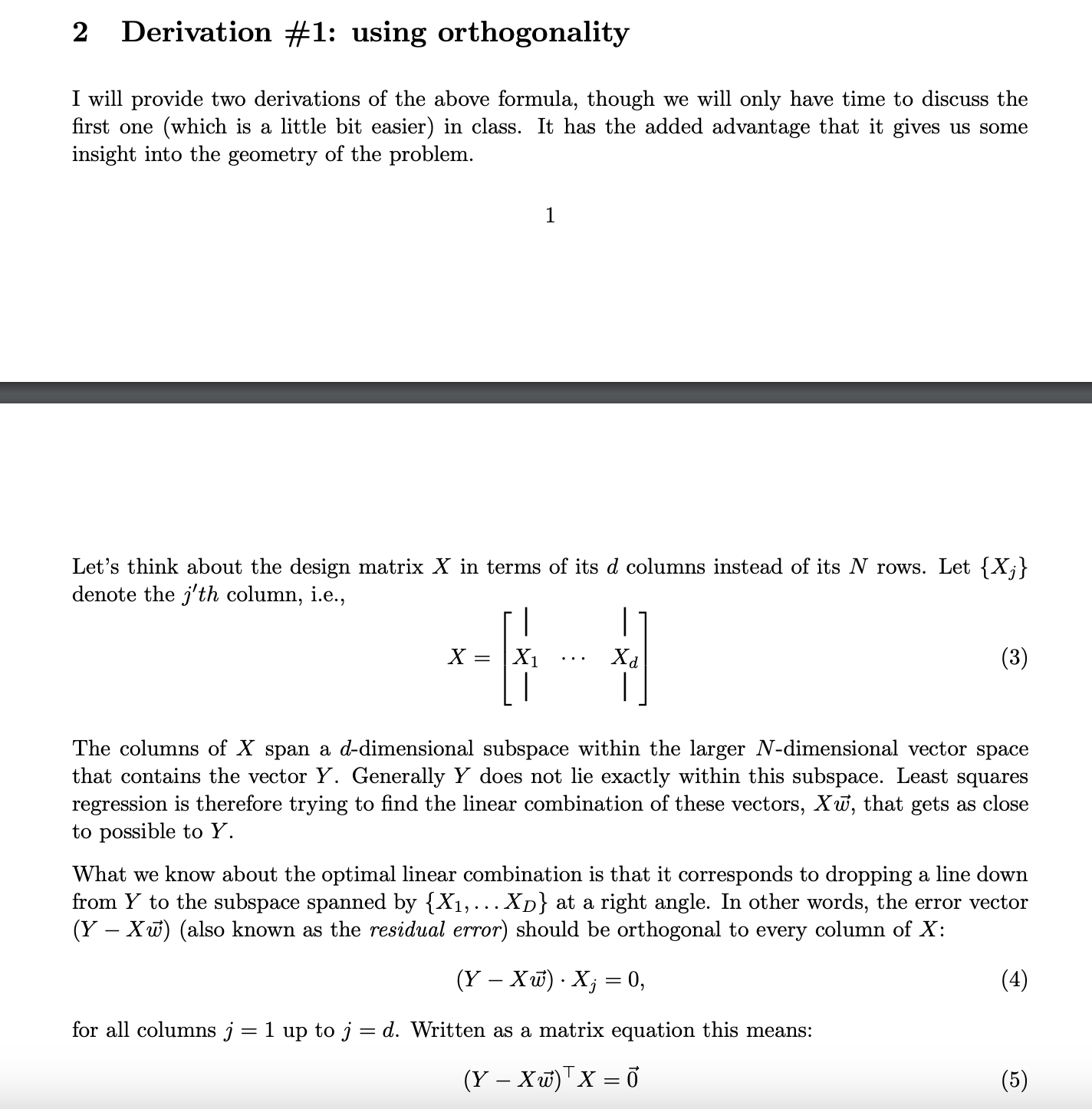
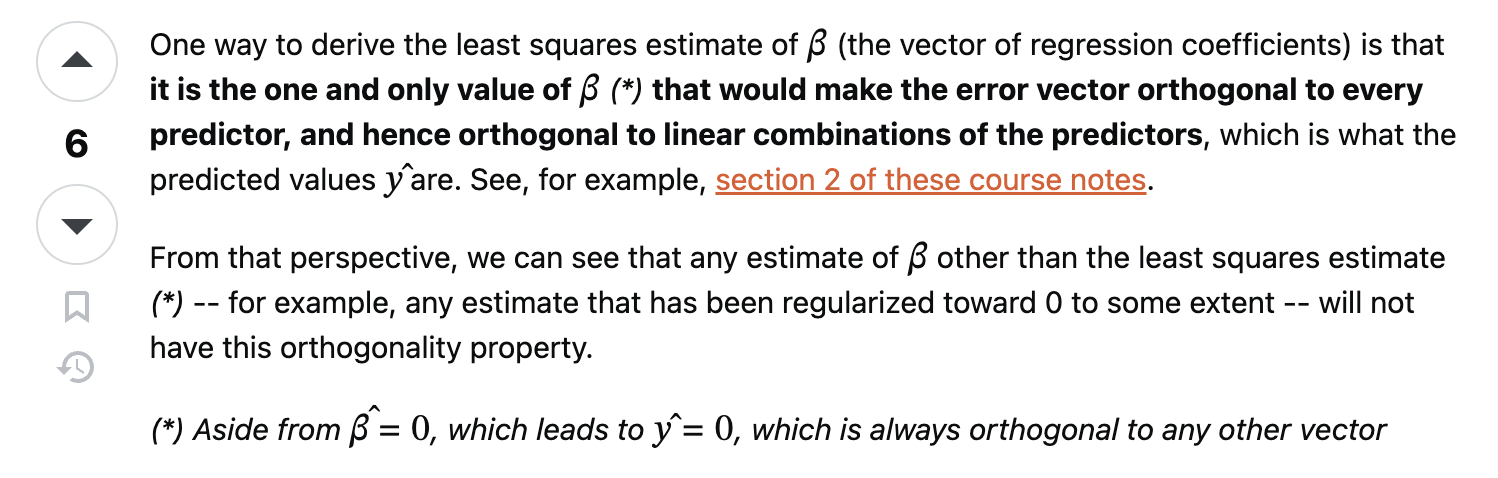
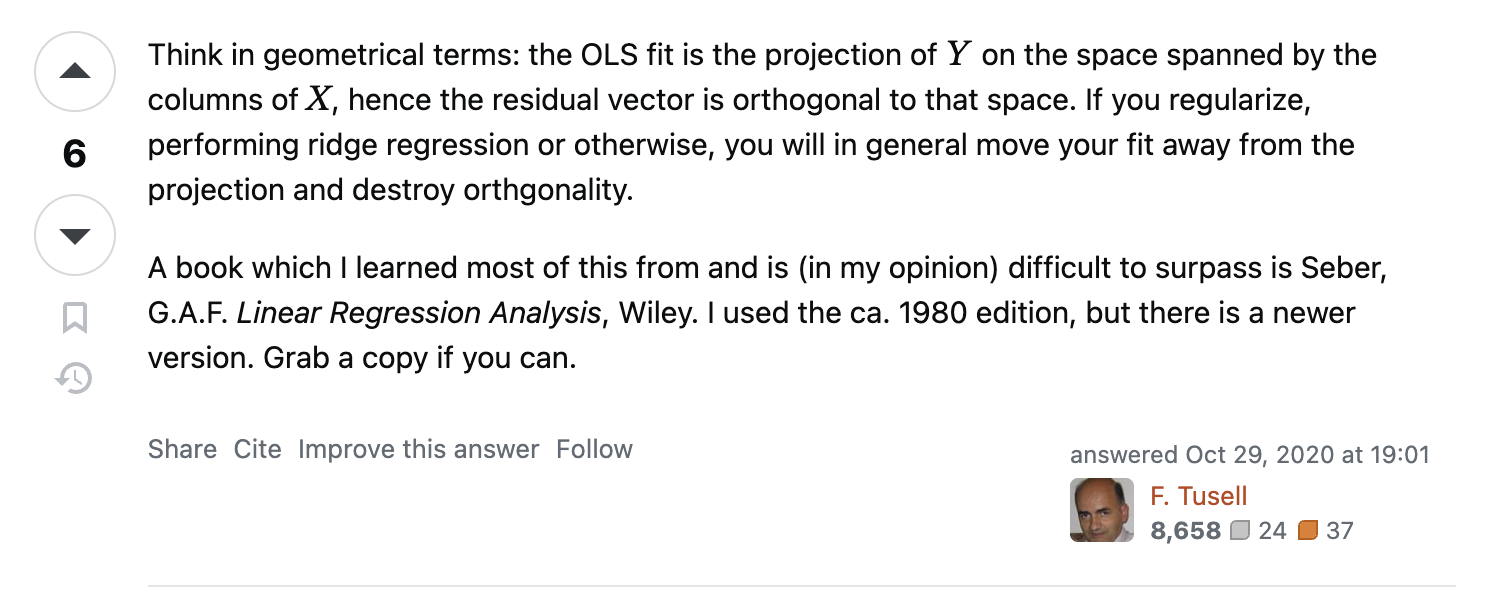
**No longer orthogonal**

**最简单的办法是，OLS的X\*B\_hat是projection matrix, orthogonal; 那ridge不再是projection matrix了，自然不再orthogonal**

<https://www.1point3acres.com/bbs/thread-925252-1-1.html>

<https://stats.stackexchange.com/questions/494274/why-does-regularization-wreck-orthogonality-of-predictions-and-residuals-in-line>





**柯西不等式这么考的：有一组数据x和y用来fit两个回归方程y = beta1 \* x, x = beta2 \* y，然后已知beta1的值问beta2的取值范围。**

经典题目，这哪里柯西了

<https://www.1point3acres.com/bbs/thread-494623-1-1.html>

**y ~ x1 的线性回归系数是b1**

**y ~ x2 的线性回归系数是b2**

**y ~ x1,x2 的线性回归系数是b1' 和 b2'**

**问题来了，b1' b2' 和 b1 b2的关系是什么?**

感觉是硬算出来的

<https://www.1point3acres.com/bbs/thread-544552-1-1.html>

第一个问题没搞清楚，后来想了下: 这一题要按照定义，就可以把关系找到。

b = (X^T X)^(-1) X^T Y .

我的结论长这个样子：

1 \* b1' + rho \* b2' = b1

rho \* b1' + 1 \* b2' = b2

这里面的rho = <x1,x2> / sqrt( < x1^2 > <x2^2> )

**Confidence Interval of Linear Regression Coefficient**

<https://www.1point3acres.com/bbs/thread-600747-1-1.html>

STAT 521 Lec 6

**问了为什‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍么OLS是minimize r^2, 而不是r^4**

(computationally cose? 不确定)

**或者|r|**

(not smooth and convex, hard to calculate gradient).

<https://www.1point3acres.com/bbs/thread-451111-1-1.html>

**linear regression的objective function, 然后如果不平方，用训练好的model算error, 是大于零小于零还是等于零。。**

（存疑，不用太当真）

答案 等于零

理论上residuals应该zero mean gaussian

不平方的话 有正有负 抵消了

<https://www.1point3acres.com/bbs/thread-474733-2-1.html>

**给了几个点，问linear regression line长啥样，加了个leverage point问长啥样，然后问如何缩小leverage point的影响。答设定一个threshold，当时说的是mean(y)+-若干sigma，每个点的error为min(threshold, residual)。**

<https://www.1point3acres.com/bbs/thread-595847-1-1.html>

小哥说可以，但我们也可以改loss function，比方说改成L1-norm，leverage point的影响就小了。然后最后又补了个方法就是regression完事儿之后找leverage point然后降低它在regression model里的权重，直接扔了也行。小哥说一般确实是这么操作的。

**(其实就是multicollinearity)regression, n个predictor, m个是剩下n-m个的linear combination, 问会有什么问题？ (answer: rank deficient, solution not unique)然后很迷的一问来了 问我solution set, what does it look like 我真的是？？？？**

然后答案是hyperplane。。小哥告诉我了。。我以为什么高难度神奇的形状。。

这题别当真，原po的知识水平和表达能力都非常差<https://www.1point3acres.com/bbs/thread-479357-1-1.html>

**得当时quant research问了一些优化问题，比如说how to invert a high dimensional covariance matrix, 怎么用三个1 dimensional的随机变量simulate 3 dimensional random variables with given covariance structure**

<https://www.1point3acres.com/bbs/thread-466148-1-1.html>

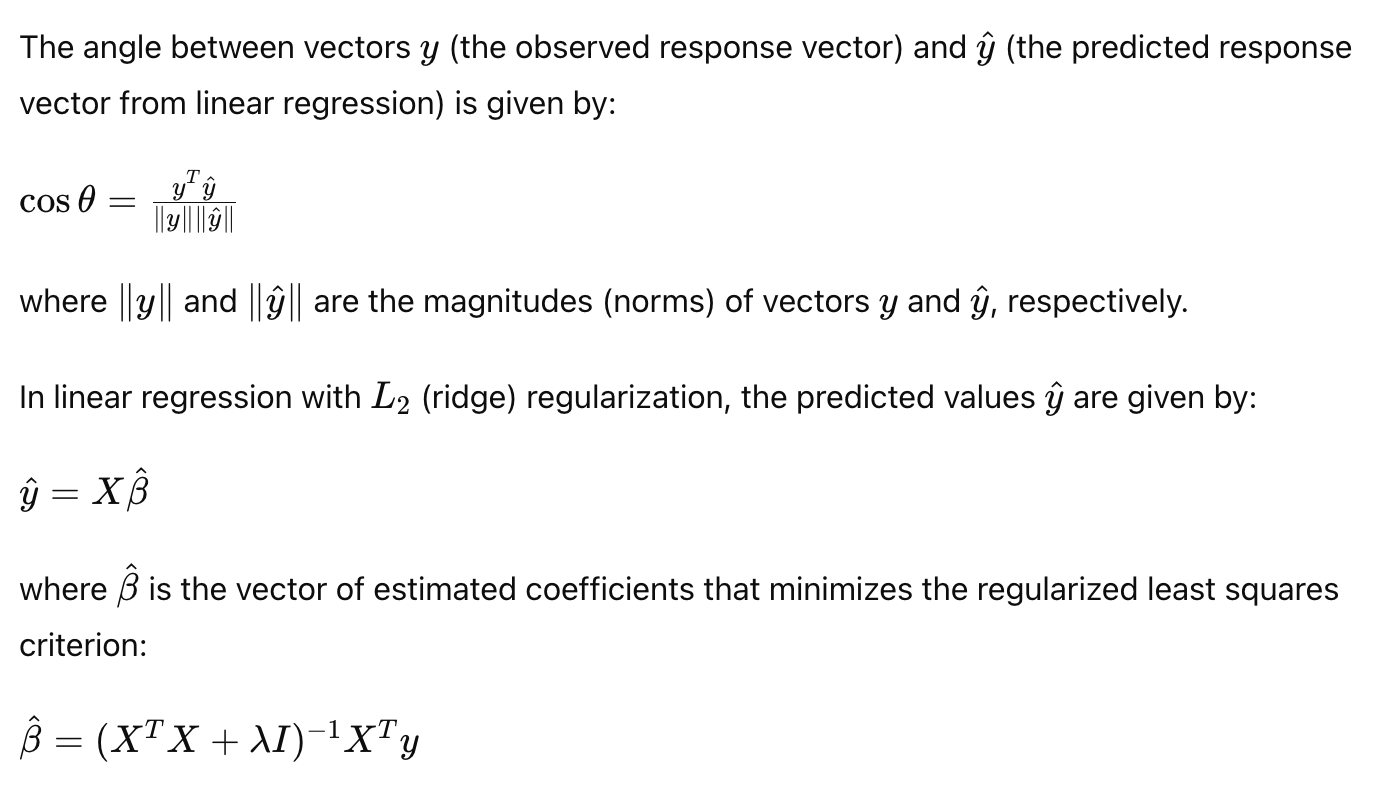
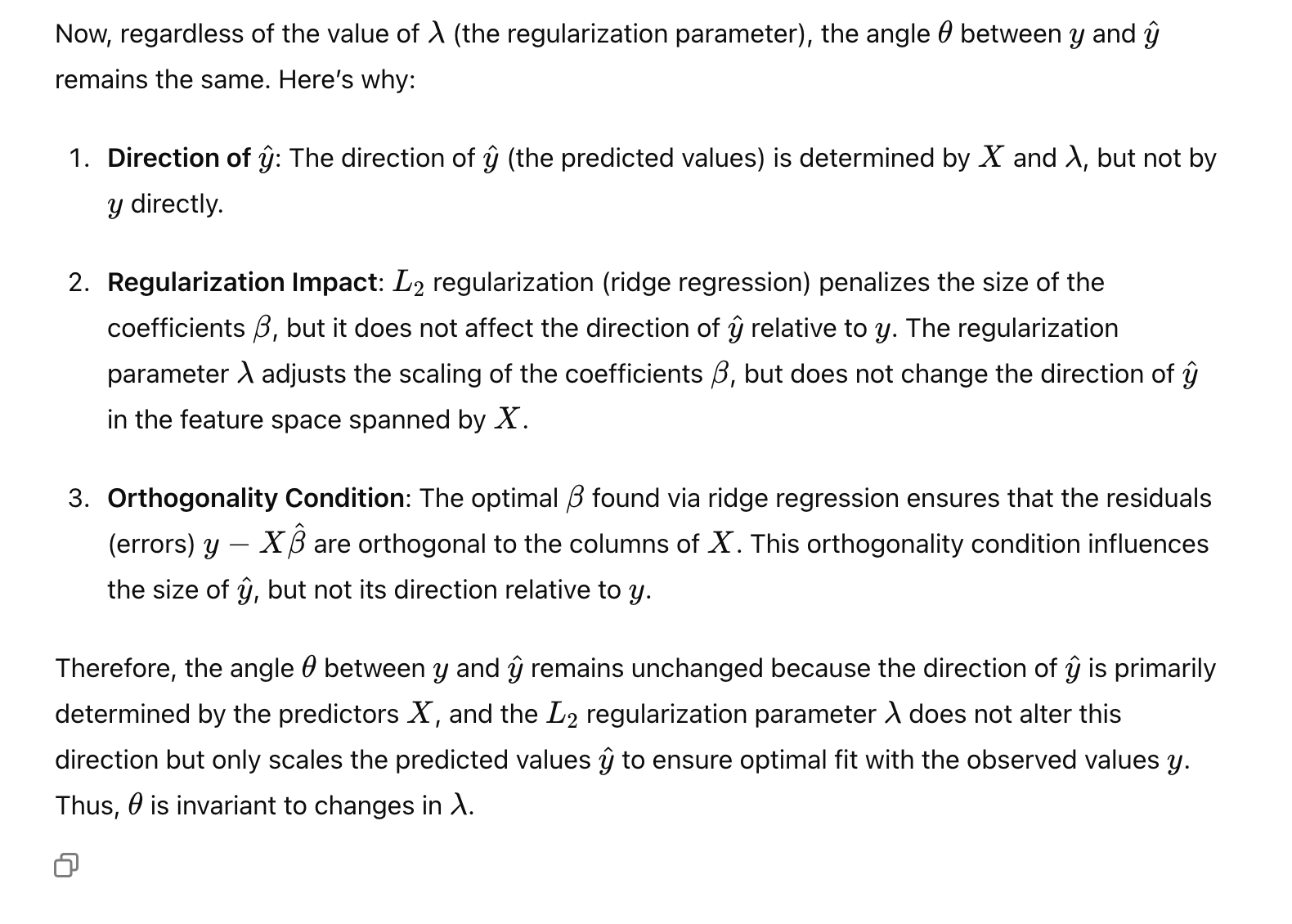
这两问都没找到合适的答案

问了chatgpt,invert那一问可能可以用cholesky；符合自己的猜想

<https://www.1point3acres.com/bbs/thread-466148-1-1.html>

**Linear regression y = X \* beta + epsilon, where X is an N \* p matrix and y is a vector of dimension N. Whatever lambda you use in the L\_2 regularization, the angle between vector y and y^hat is the same. Try to ‍‌‌‍‌‍‌‌‍‍‍‍‍‌‍‌‌‌‍‍analyze why.**

<https://www.1point3acres.com/bbs/thread-543875-1-1.html>



**LASSO solution through KKT**

这东西太nb了

<https://stats.stackexchange.com/questions/17781/derivation-of-closed-form-lasso-solution>

