IDS.131 PSet 1

Dimitris Koutentakis

September 2018

1 Salk Vaccine Field Trial Analysis

a) Comparing the two studies

The two main differences between the two groups are that the first one has the groups split according to grades and the second one is double blind. The first experiment introduces some bias because the vaccine group and the no vaccine group are split according to their grade. However, grades 1 and 3 is a more representative number because it includes those who didn't consent as well as those who did not receive the The second experiment, contrary to the first one is a double-blind experiment which means neither the subjects, nor the testers know who receives what medication. This method has less bias and is considered to be better. The second test would have more power than the first one.

b) Effectiveness mesasures

The numbers that show the effectiveness of the vaccine are

1. For the first experiment the "vaccine" and the "no vaccine" groups, and the polio rate numbers. (25 and 54 per 100,000)

2. For the second experiment, the numbers that have more importance are the "Treatment" and the "Control" groups and the "Polio Rate" numbers". (28 and 71 per 100,000)

c) Rate explanation

In the two studies, the control groups have higher rate of polio than than the no-consent groups even though neither of them received the vaccines, because of selection bias. Members of the no-consent group have chosen not to consent for a specific reason (e.g. they are healthy), so they are less likely to be sick. On the other hand, the control group is made of all sorts of people and represents the general population.

d) Bias causes

Yes, the results of the NFIP study would be biased because it not being blind introduces some selection bias, since the subjects decide themselves if they want the vaccine or not. Furthermore, since the disease is infectious, if one person in a grade gets sick more people can easily get sick in the same grade which can play a role in larger schools.

e) Analyzing parents' behavior

No, the parents were most probably not right. As discussed above, the difference can be attributed to selection bias. This means that those who did not consent were probably low risk individuals with low chance of getting polio anyways. In contrast those who did consent, did so for a reason, (e.g. being high-risk). However the students of the following year are members of the general population, and making decisions as if being part of the low-risk group is not the right way to act on this information.

2 NASA Compton Gamma Ray Observatory

a) Potential data model

A plausible model for this process is the following:
Let $y_1, y_2, ..., y_{100}$ be the gamma ray counts in the time intervals $t_1, t_2, ..., t_{100}$. We suppose that in each interval we have a poisson process with parameter λ_i. Then:

$$P(y_i) = \frac{(\lambda_i t_i)^{y_i} e^{-\lambda_i}}{y_i!}$$

b) Null and alternative hypothesis

The null hypothesis is:

$$H_0 : \lambda_1 = \lambda_2 = ... = \lambda_{100} = \lambda$$

. The alternative hypothesis (H_0) is that not all λ_i are equal.

c) Estimator calculation for null model

The most plausible parameter value for the model described above under the null hypothesis, is the MLE for λ, which is as follows:

$$l(\lambda) = P(y_1, y_2, ...y_{100}, \lambda) = \prod_{i=1}^{100} \Big(\frac{(\lambda t_i)^{y_i} e^{-\lambda t_i}}{y_i!} \Big)$$

$$\Rightarrow log(l(\lambda)) = \sum_{i=1}^{100} \Big(log\Big(\frac{(\lambda t_i)^{y_i} e^{-\lambda t_i}}{y_i!} \Big) \Big)$$

$$= \sum_{i=1}^{100} \Big(y_i log(\lambda t_i) - \lambda t_i - log(y_i!) \Big), \text{Thus:}$$

$$\frac{\partial log(l(\lambda))}{\partial \lambda} = \sum_{i=1}^{100} \Big(\frac{y_i}{\lambda} - t_i \Big)$$

$$= \frac{1}{\lambda} \sum_{i=1}^{100} y_i - \sum_{i=1}^{100} t_i$$

By setting $\frac{\partial log(l(\lambda))}{\partial \lambda} = 0$, we get:

$$\hat{\lambda} = \frac{\sum_{i=1}^{100} y_i}{\sum_{i=1}^{100} t_i}$$

d) Estimator calculation for alternative model

For the alternative model (H_A), the most plausible parameter values are the MLEs for each of the intervals. We already know the MLE for a poisson distribution. Thus:

$$\hat{\lambda_i} = \frac{y_i}{t_i}$$

e) Test statistic for H_0

The test statistic for this model will be:

$$\Lambda = -2log\Big(\frac{P(y_1, y_2, ..., y_n | H_0)}{P(y_1, y_2, ..., y_n | H_0 \cup H_A)} \Big)$$

$$= -2\Big(\sum_{i=0}^{100} (y_i log(\lambda t_i) - \lambda t_i - log(y_i!)) - \sum_{i=0}^{100} (y_i log(\lambda_i t_i) - \lambda_i t_i - log(y_i!)) \Big)$$

$$= 2\sum_{i=0}^{100} (-y_i log(\lambda t_i) + \lambda t_i) + 2\sum_{i=0}^{100} (y_i log(\lambda_i t_i) - \lambda_i t_i)$$

$$\Lambda \sim \chi^2_{dim(H_0 \cup H_A) - dim(H_0)}$$

Under the null, Λ will have a χ^2 distribution with $N-1 = 99$ degrees of freedom. That is, the distribution looks like the graph in figure 1.

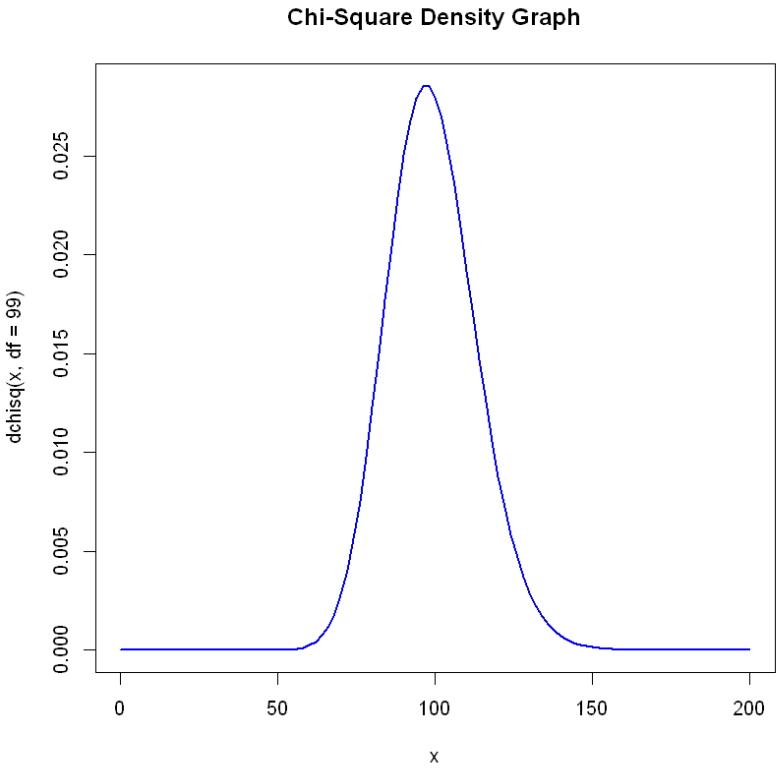


Figure 1: Chi Squared distribution with 99 degrees of freedom

f) Rejection region for $\alpha = 0.05$

With the function qchisq(0.05, 99) in `R`, we get that 77.0463 has a CDF of $\alpha = 0.05$. Thus the rejection region is the red area in figure 2

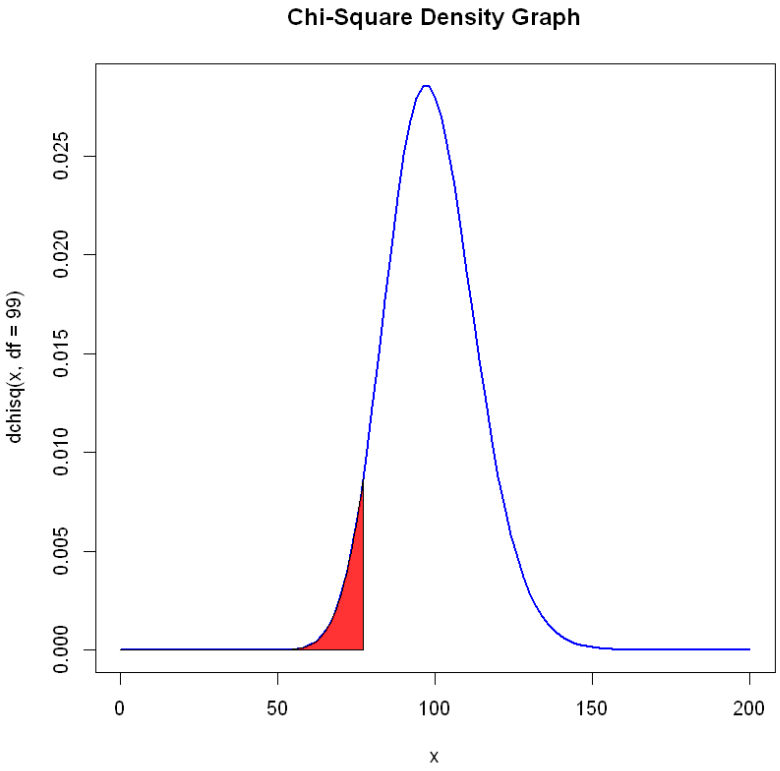


Figure 2: Rejection region

g) P-value and final conclusion

After calculating Λ in R, we found a value of roughly 104, leads to a P-value of 0.664 which means our null hypothesis is not rejected. The value of our test statistic is the red line in graph 3. Thus we can say that the emission rate is in fact constant.

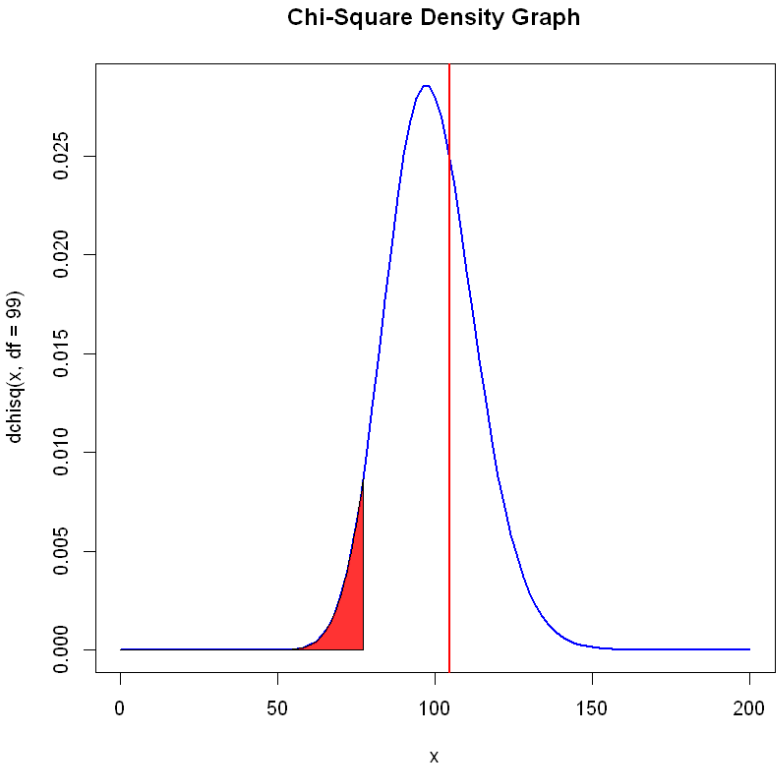


Figure 3: Chi Squared distribution with 99 degrees of freedom

3 P-Values

According to the ASA , P-values are commonly misused and misinterpreted in scientific papers. Additionally, P-values are susceptible to P-hacking, a practice in which researches try performing many tests and/or selecting the data such that they get impressive or novel results. P-values are not a good measure of evidence regarding a hypothesis, but rather a measure of the incompatibility of the data with the null hypothesis. For those reasons, banning P-values could be beneficial for the statistical community.

However, P-values are a powerful tool and give valuable information when we need to reject a hypothesis. P-values can be used to show the incompatibility of data with a hypothesis, and depending on the significance level (α), we can get trustworthy results. Especially with the combination of other metrics, one can get good insights. Instead of banning P-values there should be a more systematic approach and guidelines for the correct use of P-values, as well as better peer review for the scientific papers.

4 Detecting Leukemia types

In order to determine how many genes are associated with the different tumor types, I loaded the data in R, and performed a two-sided location test for each gene. In specific, I used the `t.test()` function in R with the two groups (first 27 and last 11) in order to compute the P-values of the null H_0 that each of the genes has no difference. Then I tested those P-values against a significance level of $\alpha = 0.05$ and then repeated the same after having corrected the P-values with the Holm-Bonferroni and the Benjamin-Hochberg corrections. The results are summarized below:

	Uncorrected P-Value	Benjamin-Hochberg	Holm-Bonferroni
Genes associated	1078	695	103

5 Why most published research findings are false

In his paper, Ioannidis explains why it is true that most published research is not true, but is in fact false. He goes through the calculations and shows how factors ranging from bias, to number of people working on a relationship, to researcher incentives and number of relationships tested tend to make it more likely that a relationship is proven to be true even if it actually is not.

The most important lesson I learned is that published papers are not to be trusted blindly. In contrast there needs to be a lot more caution when trying to prove relationships and replication.

The computations going into table 1 are basically just an application of Bayes rule, where c is the total number of researchers testing a correlation, $1 - \beta$ is the probability that the research is proven true given it is in fact true and $\frac{R}{R+1}$ is the prior probability that the relationship is true. In table 2, Ioannidis adds the bias factor, and shows how that further decreases the probability that a research is true. Finally, in table 3, he accounts for multiple studies which tend to decrease the probability even more. Ioannidis gets the result that a research is more likely to be true than false if $(1 - \beta)R > \alpha$, just by setting $PPV > \frac{1}{2}$. Then by solving, we get the relationship $(1 - \beta)R > \alpha$. This means that published research many times is not true and that we should not blindly trust it. Statisticians need to be more strict with such papers and people should try to be more careful when reading or publishing papers.

6 Regression and Gradient Descent

a) Computing OLS estimator $\hat{\beta}$ by matrix inversion

After importing the the data in `R`, I computed the β value by matrix inversion according to `beta_actual = solve(t(X)%*%X)%*%t(X)%*%y}`. The value

I found for β is:

$$\beta = \begin{bmatrix} 1.448439 \\ -4.751863 \end{bmatrix}$$

b) Implementing Gradient Descent

I implemented the gradient descent algorithm in R, as shown below:

```
gradientDesc <- function(x, y, step_size, n_iter, beta_0) {
n = nrow(y)
beta_list <- matrix(nrow=2, ncol=n_iter)
MSE_list <- matrix(ncol=n_iter)
beta_t=beta_0
iterations = 1
while(iterations <= n_iter) {
    beta_list[1:2,iterations]=beta_t
    MSE <- sum((y - x%*%beta_t) ^ 2) / n
    beta_t <- beta_t + step_size * 2*t(x)%*%(y-x %*% beta_t)
    MSE_list[iterations]=MSE
    iterations = iterations + 1
}
results <- list("beta" = beta_t, "betas" = beta_list,
"MSEs" = MSE_list, "N" = n_iter)
return(results)
}
```

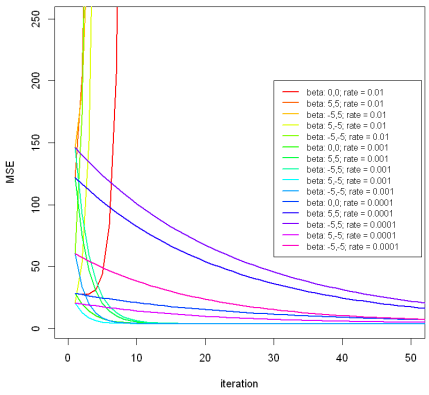
After implementing the function, I ran it several times with values of:

$$\beta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} -5 \\ -5 \end{bmatrix}, \begin{bmatrix} -5 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 \\ -5 \end{bmatrix}$$

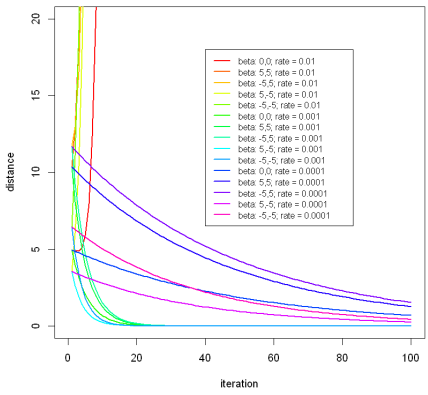
I also tried values of step size: $\eta = 0.01, 0.001, 0.0001$. The results can be seen in the following two plots, (a) of the Mean Squared Errors (MSEs) and (b) of the Distances vs the number of the iteration.

With regards to the step size we can see that independent of initialization, $\eta = 0.01$ does not converge, $\eta = 0.001$ converges quickly, and $\eta = 0.0001$ converges, but slowly (although we can expect potentially greater accuracy).

With regards to the initialization of the vector β, we can see that $\beta = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$ is the closest, and $\beta = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$ is the one with the worst error. This is expected, as the error is larger for the β further away of the actual one.



(a) MSEs vs iterations



(b) Distances vs iterations

c) Getting an overview of the data

In order to find which cities stand out, I imported all of the data into R and produced histograms for all of them. The histograms are summarized in figure 5

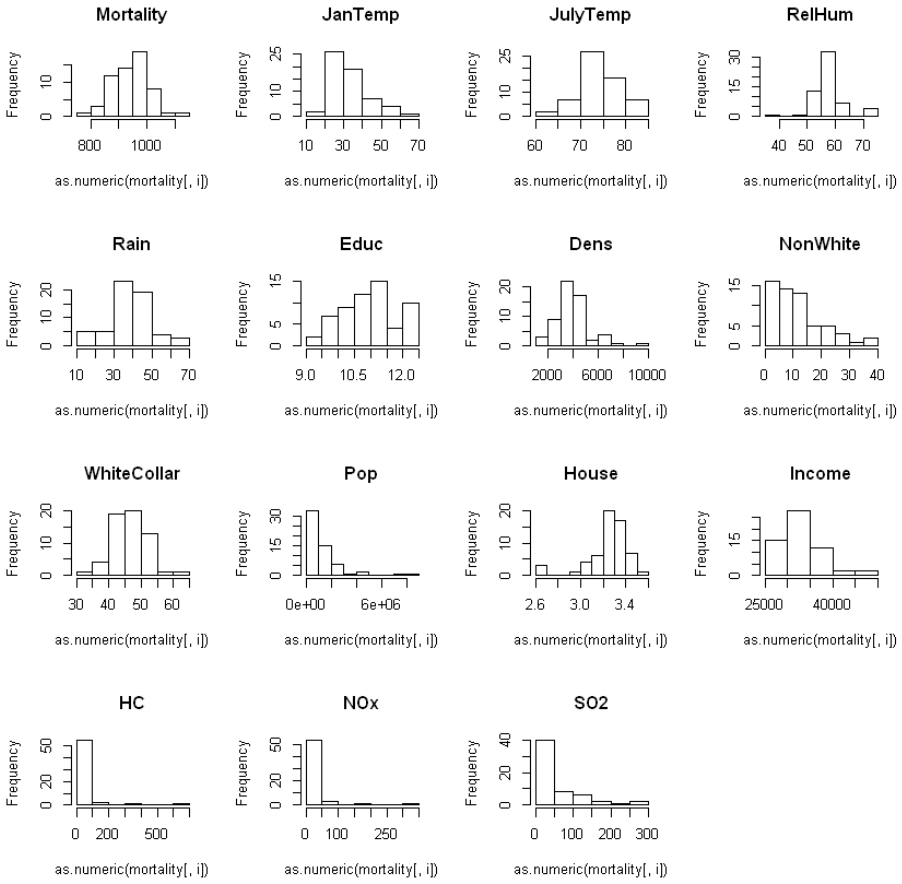


Figure 5: Histograms of each columns of the data

When comparing each of the cities to the histograms, we can see that the cities that stand out are: Los Angeles with a very low relative humidity, Allentown and Birmingham in terms of non-white people, York in white collar people and Washington DC in population. However things are even more drastically different when it comes to HC, NOx and SO2. Los Angeles stands out with very high HC and NOx levels and Pittsburg and Chicago stand out with very high SO2 levels.

It seems that the variables with the highest need to be transformed are the variables `'NonWhite', 'Pop', 'HC', 'NOx', 'SO2'`

A possible problem could be that the values of each of the variables have different orders of magnitude, and so they might be weighted differently.

d) transforming the data

In order to transform the data, we apply basic transformations to try to bring the variables back to a normal distributions. Some of these are: $log(X)$, $\frac{1}{X}$, $\sqrt{(X)}$, X^2. Then we transform their magnitudes so that every variable is between 0 and 1. The transformed histograms are shown in Figure 6

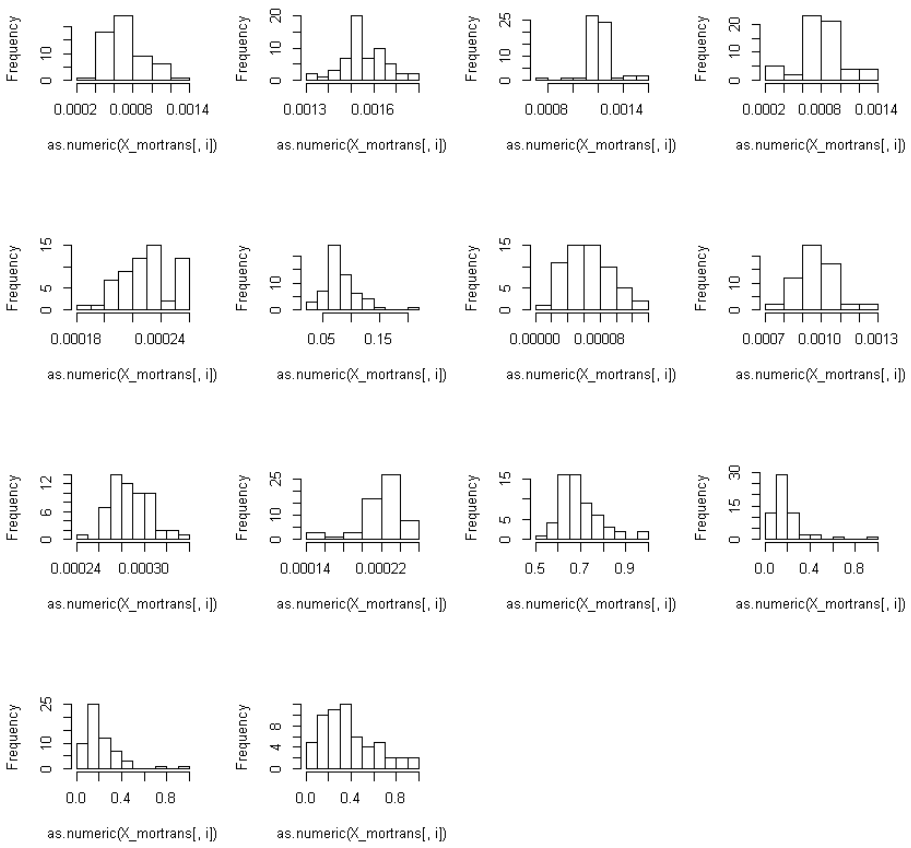


Figure 6: Histograms of the transformed data

Before the transformations, the gradient descent algorithm diverges even for very small values of the stepsize. However after the transformations it seems to converge better to: [0.12709256, 0.25274884, 0.17805350, 0.20950093, 0.02565729, 1.11810787, 0.02013619, 0.08885623, 0.03534617, 0.04160429, 1.05158450, -0.70868814, 0.37559486, 0.17441938].

7 Computational Aspects of Regression

a) Large feature matrix challenges

Storing such large matrices in memory and performing operations on them is almost impossible on regular personal computers. In specific, we would need $10 \cdot 10^7 \cdot 2 \cdot 10^2 \cdot 64$ bits, which evaluates to 160 Gigabytes of memory. Most recent personal computers have between 8 and 16 Gigabytes of memory for comparison. Another problem to consider is matrix multiplication and inversion which take up to $O(n^3)$ time. For this large matrices this would take a long time.

b) Computing regression coefficients

In order to compute the linear regression coefficients, we would need to drastically lower the number of parameters we use at every iteration. A good way to do that is with Stochastic Gradient Descent. Stochastic Gradient Descent is much faster and needs less memory than regular gradient descent. However, it is more erratic, as it usually bounces around the right answer (it is correct in expectation).

c) Gradient descent viability

We will not be able to run gradient descent, since it will not converge. The rank of the matrix $A = X^T X$ will be less than or equal to the small dimension of the matrix X, which is smaller than the dimension of A. Thus matrix A is invertible. Since it is invertible, we cannot find a solution and the gradient descent will not be able to converge somewhere.