Marlond Criollo Collaborators: Office Hours

Problem 1: The Salk Vaccine Field Trial

The first polio epidemic hit the United States in 1916. By the 1950s several vaccines against the disease had been discovered. The one developed by Jonas Salk seemed the most promising in laboratory trials. By 1954, the National Foundation for Infantile Paralysis (NFIP) was ready to try the vaccine in the real world. They ran a controlled experiment to analyze the effectiveness of the vaccine. The data is shown in the table below (grade refers to educational stage). The experiment was later repeated as a randomized controlled double-blind experiment. This data is shown in the second table below.

NFIP study		
	Size	Polio rate per 100'000
Grade 2 (vaccine)	225000	25
Grades 1 and 3 (no vaccine)	725000	54
Grade 2 (no consent)	125000	44

Randomized controlled double-blind experiment		
	Size	Polio rate per 100'000
Treatment (vaccine)	200000	28
Control (salt injection)	200000	71
No consent	350000	46

Part A

Compare the two studies and comment on the differences.

Each study choose a different way of separating subjects. The first study separated by grade and which is highly correlated with age. This can impact resistance levels. The second study created groups randomly. Another thing to note is that the second study was double-blind while the the first was not. This can potentially make our data from the first study more noisy. The NFIP study may have some bias since parents choose to get the vaccine for their kid. The second study is much more straightforward when considering potential biases and the results.

Part B

Which numbers show the effectiveness of the vaccine?

To get a sense of the effectiveness we would have to consider the Polio rate of the treatment group and the control group of the second study. Based on initial

observation it appears fairly effective.

Part C

In the two studies neither the control groups nor the no-consent groups got the vaccine. Yet the no-consent groups had a lower rate of polio. Why?

This might because of a selection bias, where parents are more cautious with their kids. They might believe the vaccine will expose their kid to the disease and this kind of attitude might permeate through the overall exposure the no-consent kid.

Part D

Polio is an infectious disease. The NFIP study was not done blind. Could this bias the results?

This could definitely bias the results as kids who knowingly got the vaccine might have been more ready to expose themselves to the disease, while in the no-consent case these families might be more cautious with the kids.

Part E

In the randomized controlled trial the children whose parents refused to participate in the trial got polio at the rate of 46 per 100,000. On the other hand, the children whose parents consented to participate got polio at a slighter higher rate of 49 per 100'000 (treatment group and control group taken together). On the basis of these numbers, in the following year some parents refused to allow their children to participate in the experiment and be exposed to this higher risk of polio. Were they right?

I do not believe they were right. Although there was a correlation that does not imply causation. The action of giving consent or not giving consent could have had these direct effects on the rate of polio in their children. This is because this was not a random action, it was based on the parents initial impression on how dangerous the disease is before and after getting a vaccine or not. The case could be that parents who gave consent may have children who were at higher risk.

Problem 2: NASA Compton Gamma Ray Observatory Data

The file gamma-ray contains a small quantity of data collected from the Compton Gamma Ray Observatory, a satellite launched by NASA in 1991 (http://cossc.gsfc.nasa.gov/). For each of 100 sequential time intervals of variable lengths (given in seconds), the number of gamma rays originating in a

particular area of the sky was recorded. You would like to check the assumption that the emission rate is constant.

Part A

What is a good model for such data? Explain your answer.

Since we want to check the assumption that emission rate is constant and we know that the number of gamma rays is discrete and non-negative we could use a Poisson distribution. If we go with this, we could have variable R_i represent the number of rays within time interval i.

Part B

Describe the null hypothesis H_0 and the alternative H_A.

Lets consider λ_i as the emission rate during interval i.

$$H_0 : \lambda = \lambda_i \forall_i$$

$$H_A : \exists i, j \text{ such that } \lambda_i \neq \lambda_j$$

Part C

What is(are) the most plausible parameter value(s) for the null model given the observations? Calculate the maximum likelihood estimate(s) (MLE) of the parameter(s). Compute the estimator(s) for these parameter(s) from the data and report the resulting value(s).

If we assume that the number of emissions are independent for each given interval. The likelihood function is equal to the product of their probability mass functions:

$$f(R) = \prod_{i=1}^{n} \frac{e^{-\lambda t_i}(\lambda t_i)^{R_i}}{R_i!}$$

Once we take the log of the function we will get

$$\ln(f(R)) = -\lambda \sum_{i=1}^{100} t_i + \ln(\lambda) \sum_{i=1}^{100} R_i + \ln\Big(\prod_{i=1}^{100} t_i^{R_i}\Big) - \Big(\prod_{i=1}^{100} R_i!\Big)$$

$$\frac{d}{d\lambda}\left(\ln(f(R)) = -\lambda\sum_{i=1}^{100} t_i + \ln(\lambda)\sum_{i=1}^{100} R_i + \ln\Big(\prod_{i=1}^{100} t_i^{R_i}\Big) - \Big(\prod_{i=1}^{100} R_i!\Big)\right)$$

$$0 = -\sum_{i=i}^{100} t_i + \frac{\sum_{i=1}^{100} R_i}{\lambda}$$

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

$$\lambda = 0.003880851$$

Part D

What is(are) the most plausible parameter value(s) for the alternative model given the observations? Calculate the MLE(s). Compute the estimator(s) for the parameter(s) from the data (you do not need to provide the value(s)).

The alternative model would have n parameters because under the alternative hypothesis we would have different λ parameters for each interval but the overall process would remain our log-likelyhood would look like this

$$\ln(f(R)) = \sum_{i=1}^{100} -\lambda t_i + \sum_{i=1}^{100} \ln(\lambda) R_i + \ln\Big(\prod_{i=1}^{100} t_i^{R_i}\Big) - \Big(\prod_{i=1}^{100} R_i!\Big)$$

With our new formula we would have to take the derivative again with respect to a single λ_i

$$0 = -t_i + \frac{R_i}{\lambda_i}$$

$$\dot{\lambda_i} = \frac{-R_i}{t_i}$$

This would be able to be calculated for each λ_i fairly easily.

Part E

Define a test statistic and plot its distribution under H0.

A test statistic we could use would be the log likelihood ratio statistic. Using Wilks theorem we can say that as our number n approaches infinity our $\Lambda(x) \rightarrow \chi_d^2$

Part F

Determine the rejection region at a significance level of 0.05. Depict it in the previous plot.

The rejection region is about 123.2252 it is the single black vertical line

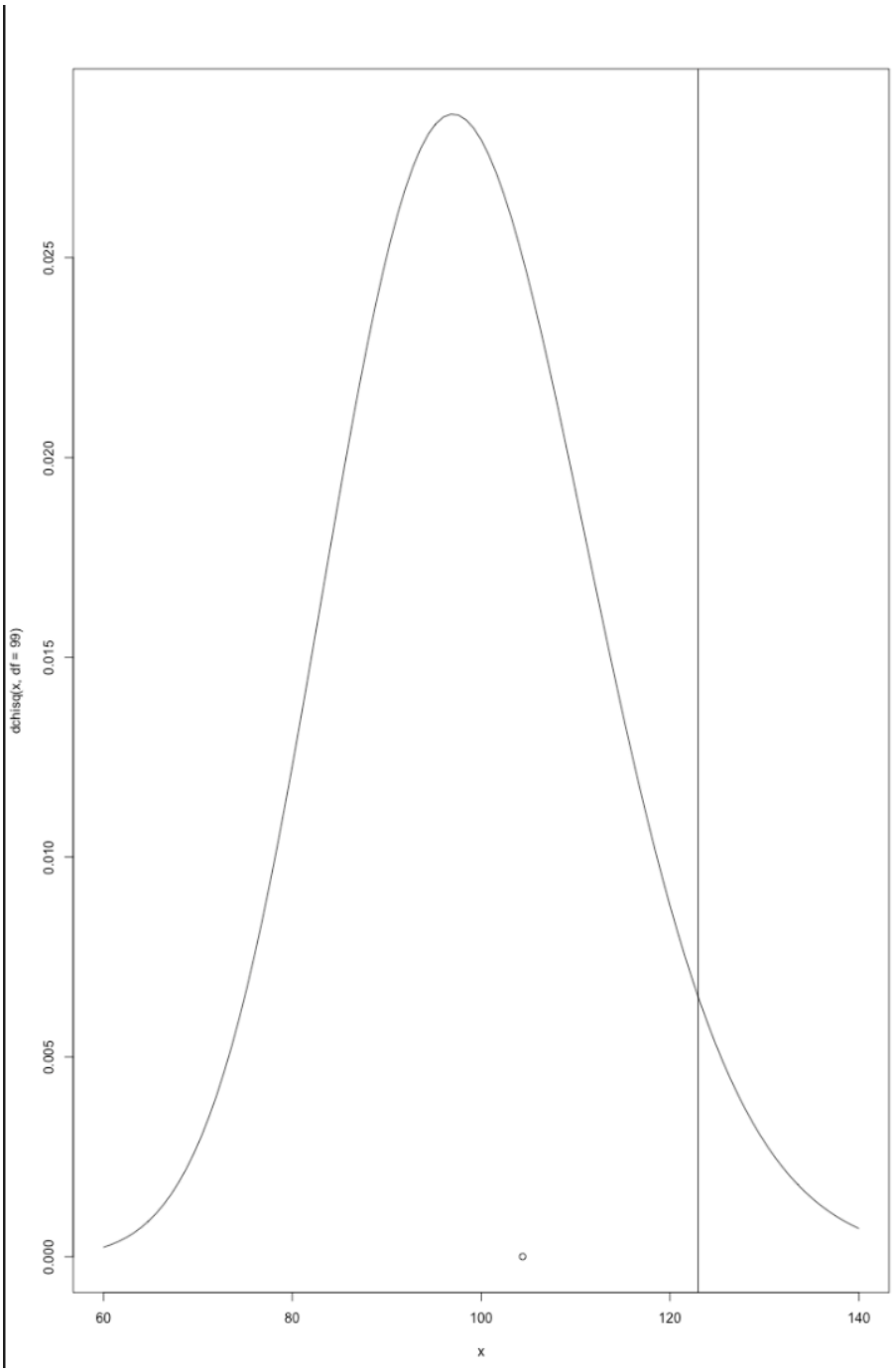


Figure 1: Screenshot 2023-02-22 at 9.01.15 PM

Part G

Also show the value of the test statistic in the previous plot. What is its p-value? Does the emission rate appear to be constant?

The value of the test statistic is 104.39. It is the black circle p value is 0.3357672 We cannot reject our null hypothesis and as a result it appears that emission rate is constant.

Problem 3

Read the statement by the American Statistical Association about p-values (Wasserstein and Lazar: The ASA's statement on p-values: context, process, and purpose) and respond to the following scenarios

Part A

A friend looking at your notes from the first lecture saw that there is a p-value of 0.0012 for the HIP study. They ask you, does that mean there's a 99.88% chance that offering a mammography decreases the risk of death from breast cancer? Explain to your friend exactly what this p-value means, including any assumptions that were made.

This isn't exactly what a p-value signifies. In this case the p-value suggests that the null hypothesis H_0 can be rejected since it less than our choosen α value of p= .05. P does not mean probability.

Part B

An economist collects data on many nation-wise variables and surprisingly finds that if she runs a regression between chocolate consumption and number of Nobel prize laureates, the coefficient is statistically significant. Should she conclude that there exists a relationship between Nobel prize and chocolate consumption?

Correlation is a statistical approach that determines how closely two variables are related and change together. This is not an explanation of the link but just shows that the relationship exists. This is an example of correlation being confused with causation. It is just like how everyone who drinks water is strongly correlated with death. Just because two things are related does not mean one causes the other. There could be another hidden factor we are not taking into account which could create this effect.Therefore she should not conclude the relationship without potential further examination and testing.

Part C

Your lab collects individual-level data on 50000 humans for 100 features, including IQ and chocolate consumption. They find that their initial hypothesis

about the relation between chocolate consumption and IQ has a p-value higher than 0.05. However, they find that there are other variables in the data set that have p-value less than 0.05, namely, a subject's family income and number of siblings. They therefore decide to not write about chocolate consumption, but rather, report these statistically significant results in their paper, and provide possible explanations. Is this sound scientific practice?

This does not seem like the correct and sound scientific practice. There is a clear lack of reporting all data and transparency with the scientific community if this inference is published. The team would be cherry-picking examples from their data. By hand selecting significant values and not reporting others, disillusions readers from understanding the researchers choices and rational for the study. This also helps create the current situation where a lot of ire has been garnered toward p-values and how easy it seems that they can be misconstrued by researchers. The lack of understanding of analysis rational, and how those analyses were deemed valid for publications, harms our ability to draw scientific conclusions

Part D

A neuroscience lab lab runs a randomized experiment on 100 mice by adding chocolate in half of the mice's diet and another food of the equivalent calories in another half's diet. They find that the difference between the two groups' time in solving a maze puzzle has p-value lower than 0.05. Should they conclude that chocolate consumption leads to improved cognitive power in mice?

They should not as they have potentially introduced confounding variables in the other food. We do not know how this inclusion of another variable could impact a mouse's ability to run through a maze nor do we know the impact chocolate might have on the mice. The p-value on its own does not provide sufficient context or evidence to draw full conclusions on an experiment or to reject the null hypothesis.

Part E

Should p-values be banned from scientific papers? Provide at least one argument for and against this proposal.

In modern society p-values are easily misconstrued, misrepresented and misunderstood. By over-emphasizing p-values outside of the context of a study can lead to very real repercussions outside of scientific experiments. For example, when a p-value is seen without context and it influences actual political decisions and policies. Banning p-values would prevent their abuse in the scientific process. A p-value is designed to improve our understanding of a given result in context of hypotheses. This in turn helps establish whether a given result is statistically significant. When done right a p-value can holistically capture the results in a single hypothesis and hypothesis testing. Despite the issues with

p-values there are steps that can be taken to rectify the state of p-values today. This could be potentially done by having researchers be more transparent with their data collection, and analysis processes.

Problem 4

The data set `golub` consists of the expression levels of 3051 genes for 38 tumor mRNA samples. Each tumor mRNA sample comes from one patient (i.e. 38 patients total), and 27 of these tumor samples correspond to acute lymphoblastic leukemia (ALL) and the remaining 11 to acute myeloid leukemia (AML).

How many genes are associated with the different tumor types (meaning that their expression level differs between the two tumor types) using

Part A

If we use a wilcoxon signed test to use uncorrected p-values we get 1055 genes

Part B

(ii) the Holm-Bonferroni correction

we end up getting 92 genes

Part C

(iii) the Benjamini-Hochberg correction? Feel free to use libraries for multiple hypothesis testing in R or python. You can use $= 0.05$ for the significance threshold.

Using this correction we get 679 genes

Problem 5

A.

i. Read in the synthetic data matrix syn_X.csv and the vector syn_y.csv of "observations". Compute the OLS estimator $\hat{\beta}$ by matrix inversion.

$\hat{\beta} =$

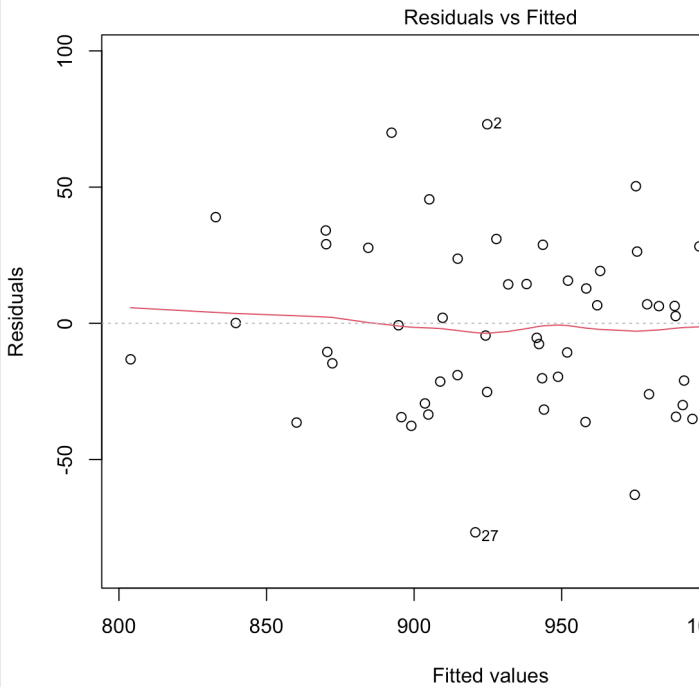
	V1
[1,]	1.263971
[2,]	-4.597993
[3,]	1.929606

Next, we look at some real data. General Motors collected data (found in mortality.csv) from 60 US cities to study the contribution of air pollution to mortality. The dependent variable is the age adjusted mortality (Mortality). The data include variables measuring climate characteristics (JanTemp, JulyTemp, RelHum, Rain), variables measuring demographic characteristics of the cities (Educ, Dens, NonWhite, WhiteCollar, Pop, House, Income), and variables recording the pollution potential of three different air pollutants (HC, NOx, SO2).

ii. Get an overview of the data and account for possible problems. Which cities stand out? Which of the variables need to be transformed?

There is one city that stands out with the highest NOx and HC and unsurprisingly its LA. Most variables seem fairly suitable for analysis there are not a significant amount of outliers and their means are similar thoughout. However when looking at Q-Q plots of NOx,HC, SO2, POP, NonWhite, and Dens they all seem almost exponential to varying degrees. Indicating that for analysis we likely want to apply a logarithmic transformation.

iii. Carry out a multiple linear regression containing all variables with the necessary transformations (with matrix inversion as in (i)). Does the model fit well? Check the residuals.



The transformations we need are described above.

```
Residuals:
    Min      1Q  Median      3Q     Max
-76.711 -23.293  -0.719  21.476  85.383

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   985.466953 349.348252   2.821 0.007158 **
JanTemp        -2.005738   0.930923  -2.155 0.036710 *
JulyTemp       -0.516908   2.113526  -0.245 0.807925
RelHum          0.157708   1.151171   0.137 0.891657
Rain            2.265185   0.611979   3.701 0.000594 ***
Educ          -11.945792   9.416430  -1.269 0.211250
log(Dens)       7.531643  17.570592   0.429 0.670269
log(NonWhite)  38.297479   9.237798   4.146 0.000152 ***
WhiteCollar    -1.659682   1.221688  -1.359 0.181225
log(Pop)        5.756244   8.210653   0.701 0.486951
House         -12.147347  40.011482  -0.304 0.762866
Income         -0.001104   0.001429  -0.772 0.444154
log(HC)       -14.877339  16.072456  -0.926 0.359681
log(NOx)       35.187766  15.446738   2.278 0.027638 *
log(SO2)       -4.562445   8.011445  -0.569 0.571919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.86 on 44 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.6514
F-statistic: 8.741 on 14 and 44 DF,  p-value: 1.286e-08
```

I managed to get these and the model pretty alright and the residual vs fitted graph show that the residuals are pretty low compared to the fitted values themselves.

B

In this problem, we will consider some computational challenges that arise in practice when performing linear regression.

i

Suppose you have a problem in which the data matrix, X, has 100 million rows and 200 columns (each row is a data point and each column is a feature / covariate). What challenge will arise when you try to apply the matrix inversion method to compute the regression coefficients as in the previous problem? (Hint: if each entry is a 64 bit float, how much memory will be required to store X?)

This would be pretty intensive for a computer to not only process but also store. The amount of memory needed would be on the order of 10^{11} bits to just store the matrix and let alone process the inversions of the matrix. The matrix inversion would cost on the order of $O(n^2)$ to perform, making the overall computation difficult.

ii

Suggest one method that will allow you to compute linear regression coefficients for this problem. Be specific. Discuss pros and cons of your approach.

One potential way to calculate linear regression coefficients would be to use mini-batch gradient descent. Here we would be able to sample our data set and be able to iterate throughout are sample pools as we update our parameter. A potential downside would be from the introduction of noise which would impact our parameters. This can be from the sampling process itself which can vary a lot to reduce this by increasing our number of samples to more easily have a normal distribution of points, but would also take more space and time.