False Discovery Rate and Loss

Lecture 2

Connor Dowd

April 1st, 2021

Today's Class

1. Assorted Business
    - Predictions
    - Questions
2. Quick Review
    - Regression
    - False Discovery Intro
3. False Discovery Rate, More than you wanted to know
4. Loss functions
5. My prediction walkthrough
6. Homework intro (if time?)

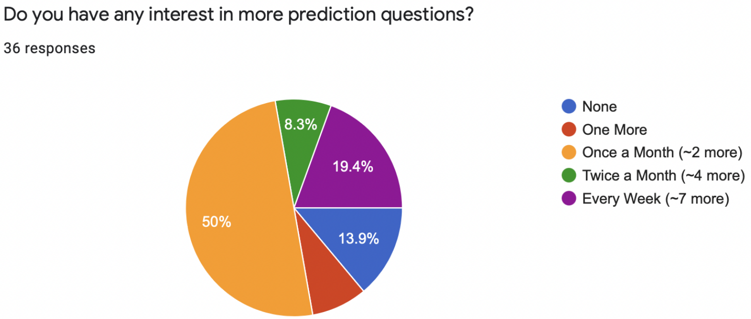
Assorted Business

Predictions

"How many people in the US will have had at least one dose by end of day on April 30th?"

- Prediction: 148 million
- 90% CI: [130,169] million.
- Based on CDC trend data – not what I gave you
  - but clearly available on the page with target numbers.
  - I pulled it in directions that felt better. Code online/later.
- You don't always have the best data
- But you could probably still do pretty well with the data I gave. CI calibration would be tough.

Interest



Do you have any interest in more prediction questions?

36 responses

Other

- R comments
    - 1-indexing
    - Usually you want to save scripts, not workspaces.
    - Stay organized. Folders for homeworks, etc.
    - Consider using shared drives or github to collaborate
- Office hours will be Fridays at 9AM

Questions from you?

Quick Review

Regression

The basic model is as follows:

$Perc.OneDose = \beta_0 + \beta_1 Delivered.100k + \beta_2 Perc.TwoDose + \epsilon$
Where $E[\epsilon] = 0$.

We care about β_1 or perhaps β_2. What are they?

Testing

We can compare pvalues, which are measure of extremity, to a pre-set threshold (α) which controls our false discovery chance.

But with lots of variables, how do we think about things?

1. No correction? $p\alpha$ false rejections
2. Bonferonni? 5% chance of any false rejections.

Both seem aggressive. Want a middle ground.

FDR Redux

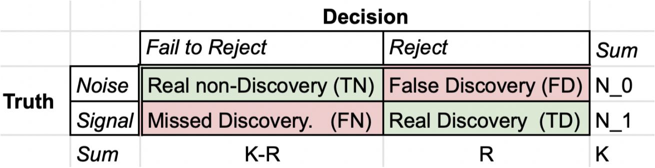
Large Scale Testing

Notation Changed

We wish to test K simultaneous null hypothesis:

$$H0_1, H0_2, \ldots, H0_K$$

Out of the K null hypothesis, N_0 are true nulls and $N_1 = K - N_0$ are false – i.e. there is an effect.



		Decision		
		Fail to Reject	*Reject*	*Sum*
Truth	*Noise*	Real non-Discovery (TN)	False Discovery (FD)	N_0
	Signal	Missed Discovery. (FN)	Real Discovery (TD)	N_1
	Sum	K-R	R	K

False Discovery Rate

FD Proportion = False positives / #Significant = $\frac{FD}{R}$
We can't know this.

We can control its expectation though: False Discovery Rate, $FDR = E[FDP]$.

If all tests are tested at α level, we have $\alpha = E[FD/N_0]$, whereas $FDR = E[FD/R]$

We can find in-sample analogues (ish) of these things.

FDR Control

Suppose we want to know that $FDR \leq q = 0.1$.

Benjamini + Hochberg Algorithm

1. Rank your p-values smallest to largest.
2. Set p-value cutoff as $\alpha^* = max\{p_{(k)} : p_{(k)} \leq q\frac{k}{K}\}$

Then $FDR \leq q$ – assuming approximate independence between tests.

Rewriting that

Step two there is a mess. Lets look at it closely.

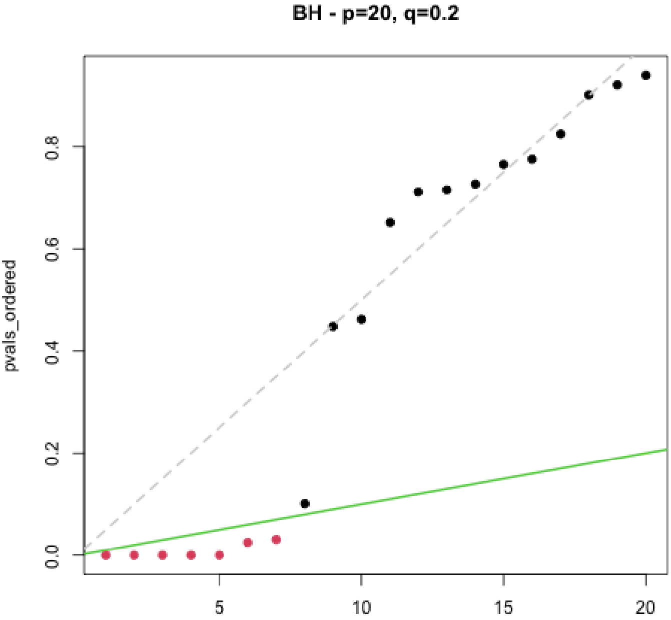
$$\alpha^* = max\{p_{(k)} : p_{(k)} \leq q\frac{k}{K}\}$$

or (because k, K are both positive)

$$\alpha^* = max\{p_{(k)} : \frac{p_{(k)}K}{k} \leq q\}$$

The secret sauce here is that $p_{(k)}K$ is the expected number of false discoveries, under the nulls, if $p_{(k)}$ were our rejection threshold. Dividing by k - the number of discoveries with that threshold - gives us an estimate of the FDR.

Understanding BH



FDR Roundup

We started with the notion that a given α, (pvalue cutoffs) can lead to a big FDR: $\alpha \to q(\alpha)$.

BH reverse that. They fix FDR, and find the relevant α. The algorithm is the key to doing that. $q \to \alpha^*(q)$

FDR is not the only way to think about these risks. But it is a very solid middle ground when we have many tests.

$=>$ Principled bounds on overall errors, while maintaining power to detect.

Example: multiple testing in GWAS

GWAS: genome-wide association studies.
Want to find genetic markers related to disease for early prevention and monitoring.

Single-nucleotide polymorphisms (SNPs) are paired DNA locations that vary across chromosomes. The allele that occurs most often is "major" (A) and the other is "minor" (a).

Question: Which ones increase risk?

Cholesterol

Willer et al, Nat Gen 2013 describe a meta-analysis of GWAS for cholesterol levels. We will focus on LDL cholesterol.

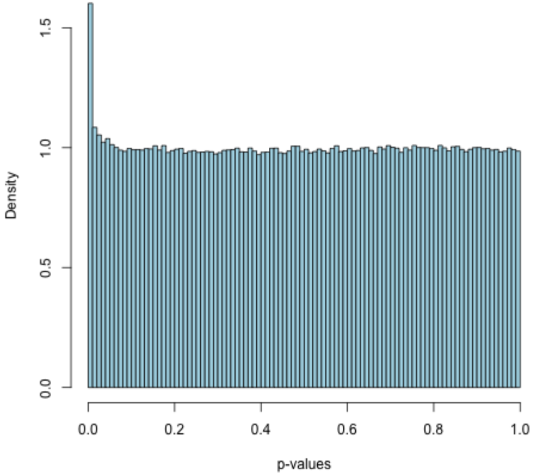
At each of 2.5 million SNPs, they fit a linear regression

$$E[LDL] = \alpha + \beta AF$$

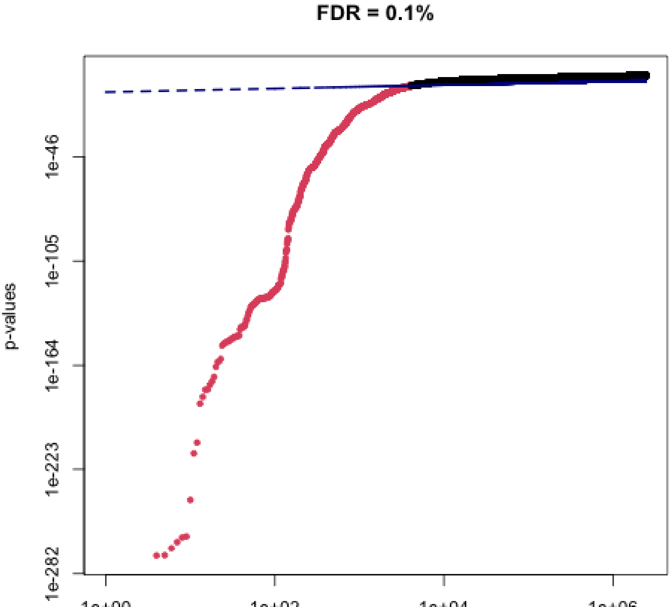
Where AF is allele frequency for the ‘trait increasing allele’.

2.5 million SNP locations.
$=>$ 2.5 million tests of $\beta = 0$
$=>$ 2.5 million p-values.

All the pvalues



BH plot (log-log)



BH Roundup

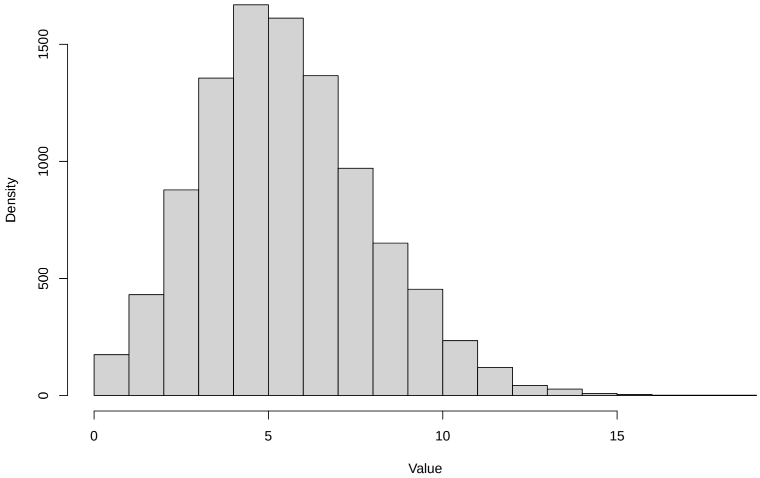
- p-values from the null distribution are uniform, and should lie along the 1/K line if there are K of them.
- FDP is the number of false discoveries divided by number of rejections. We **can't** know it.
- $$FDR = E[FDP]$$

  we can control though.
  - Fix it to be $\leq q$ for K tests
  - rank and plot p-values against rank$/K$
  - draw a line with slope q/K
  - Reject under the line.

Loss

Predictions

Suppose you have an observation coming from some distribution. What do you predict?



Some typical choices

- Mean
- Median
- Mode

Some typical choices

- Mean
- Median
- Mode

Some natural questions:

1. What if we're playing 'price is right rules' for your prediction?

Some typical choices

- Mean
- Median
- Mode

Some natural questions:

1. What if we're playing 'price is right rules' for your prediction?
2. What would motivate a choice of median over mean or mode?

Formalizing loss

At its simplest, a loss function maps the truth, and your prediction, into how unhappy you are about your error.

$$L(y, \hat{y}) = ????$$

The most common class of loss functions only cares about the magnitude of the error, not its location.

$$L(y, \hat{y}) = l(y - \hat{y}) = l(e)$$

This is not always a reasonable simplification.

Using loss

We're defining a norm against a distribution. So we need to think about all the possible values the outcome could take.

Naturally then, we're going to plug the loss into an expectation. That will let us make statements about our expected loss. (With some iffy notation)

$$E[l(e)] = E[l(y - \hat{y}] = \int l(y - \hat{y}) P[y] dy$$

This should look very familiar.

Insample

Within a sample, we can use a loss function to dictate our predictions.

We choose parameters to minimize

$$L_(Y, \hat{Y}) = L(Y, \hat{\alpha} + \hat{\beta}X)$$

l_p norm

The most common norms here are known as the l_p norms. Within sample, (and with some iffy notation) this looks like:

$$L(Y, \hat{Y}) = \left(\frac{1}{n} \sum_{i=1}^{n} |Y - \hat{Y}|^p \right)^{\frac{1}{p}}$$

Notice, we've thrown in a symmetry statement. The absolute value means that $l_p(e) = l_p(-e)$.

Again: This is not always a reasonable simplification.

Back to typical answers:

- Mean: corresponds to answer with lowest expected l_2 loss.
  - AKA: $min\sqrt{\frac{1}{n}\sum(Y-\hat{y})^2}$ is the RMSE
- Median: Answer with lowest l_1 loss
  - AKA $min\frac{1}{n}\sum|Y-\hat{Y}|$ is the MAD
- Mode: Answer with lowest l_0 loss
  - AKA $min\frac{1}{n}\sum 1(Y\neq\hat{Y})$ wants Exact predictions only

$=>$ These are different statements about how much we care about a tradeoff between infrequent large errors and frequent small errors.

How do we choose?

How many fingers do you think our dean has?

- Would you guess 10? Median/Mode?

How do we choose?

How many fingers do you think our dean has?

- Would you guess 10? Median/Mode?
- Would you guess 9.9? Mean?

How do we choose?

How many fingers do you think our dean has?

- Would you guess 10? Median/Mode?
- Would you guess 9.9? Mean?
- How do *you* choose?

How do we choose?

How many fingers do you think our dean has?

- Would you guess 10? Median/Mode?
- Would you guess 9.9? Mean?
- How do *you* choose?
- What if we were competing to be closest?

How do we choose?

How many fingers do you think our dean has?

- Would you guess 10? Median/Mode?
- Would you guess 9.9? Mean?
- How do *you* choose?
- What if we were competing to be closest?
- What if it was a random lumberjack?

Loss function last thoughts *for now*

Loss functions come from the context of a situation. No generalizable advice here.

- Standard loss functions lean on symmetry and location-indifference type assumptions that may not be reasonable.
- Very important for making actionable predictions
- And important to bake in very early
    - They drive every choice of statistic
- Price-is-right rules, competitions more generally are going to screw with this.
    - "Winners curse"

Homework Introduction

Week 1 HW

Dataset of $\sim$13k reviews for some products, collected in 2012.

Reviews include product details, ratings, and plain text comments.

We will look for words associated with good/bad ratings.

Assignment online

I will now go through the code at the start, introduce you to the datasets, run some things, and comment on various features that may help you understand R and large datasets.

Wrap up

Things to do

Before Tuesday:

- Homework

Rehash

- False Discovery Rates can be controlled
- Understanding our loss function is critical
- You have homework

Bye!