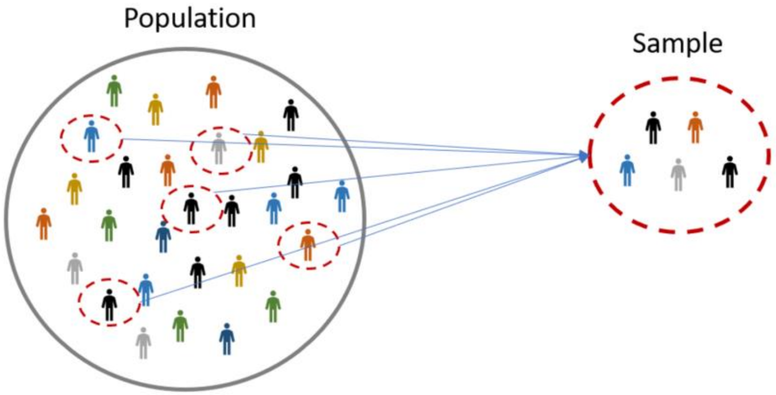
Sample Size Calculation for Logistic Regression

[redacted] Daniel Girvitz, [redacted]

Additional authors redacted 2022-06-19

Contents

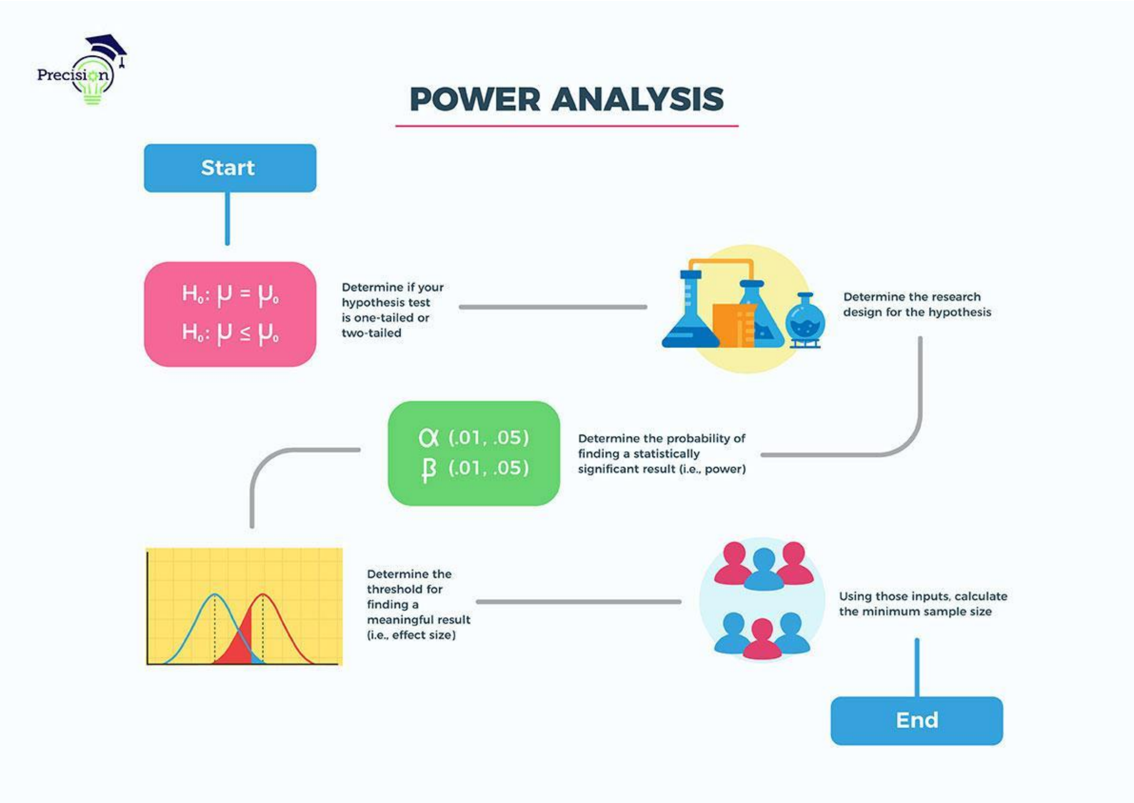
Introduction



- In a perfect world, larger sample sizes will always yield better estimates of the population parameters, but they may not be very efficient or cost effective.
- Determining sample sizes is important as we don't want a very large sample size (which will be more expensive and require more resources), or too small (which would yield less confidence in the result of the analysis).
- The goal of calculating the minimum sample size is to increase the statistical power of the experiment (decrease the probability of Type II error), while reduce the amount of resources used for the experiment.

Power Analysis

- Statistical power is the probability of detecting an effect in the experiment correctly. In other words, it is the probability of making the right decision when performing a hypothesis test.
- Power analysis allows us to determine the sample size needed for a certain effect size and certain level of confidence.
- It is comprised of <u>four</u> variables.
  - Sample size (unknown)
  - Effect size
  - Significance level
  - Power
- Thus, we can estimate the minimum sample size necessary for a given level of significance and power.



Precision

POWER ANALYSIS

Start

$H_0: \mu = \mu_0$

$H_0: \mu \leq \mu_0$

Determine if your hypothesis test is one-tailed or two-tailed

Determine the research design for the hypothesis

Determine the probability of finding a statistically significant result (i.e., power)

α (.01, .05)

β (.01, .05)

Determine the threshold for finding a meaningful result (i.e., effect size)

Using those inputs, calculate the minimum sample size

End

How to determine the sample size for logistic regression?

Concato et al. / Peduzzi et al.

- 10 is acceptable for both linear regression and cox regression

Austin / Steyerberg

- EPV = 20 instead
- Studies small to moderate sample sizes (ex. <100)

EPV

- Event per Variable (EPV)
  - EPV = # of events / d.o.f
- n = 100 + xi
  - x = EPV, i = # of independent variables

Nemes et al.

- Proved that large samples sizes (500) will increase accuracy
- Represented parameters in target population
- Derived from evaluating few populations (based on various statistical tests)
- Suggested EPV = 50
- Proved that minimum sample size of 500 will yield valid sample estimates

Whittemore (1981) / Hsieh (1998)

- Cases of Distribution: Standard Normal, Standard Exponential, Poisson, Bernoulli

$$N = \frac{(Z_\alpha + e^{-\frac{A^2}{4}}\, Z_\beta)^2}{e^{\gamma_0} A^2}$$

$$N = \frac{\left(Z_\alpha + e^{-\frac{A^2}{4}}\, Z_\beta\right)^2}{e^{\gamma_0} A^2} * \left[1 + 2^{e^{\gamma_0}} * \frac{[1 + (1 + A^2)\, e^{\frac{5A^2}{4}}]}{(1 + 2\, e^{\gamma_0}}\right]$$

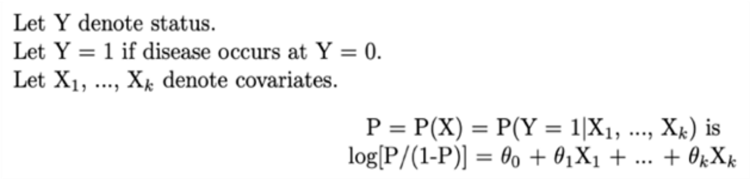
$$\mathbf{N_M = N_1/(1 - \rho^2).}$$

- Small response probability

- Assumption
  - Conditional distributions of X|Y = 1 and X|Y = 0 have normal distributions with equal variance under Ha

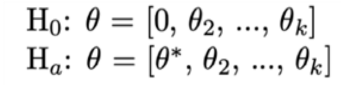
$$N = \frac{\left\{Z_\alpha \sqrt{\frac{P(1-P)}{\pi}} + Z_\beta \sqrt{[P_1(1 - P_1) + P_2(1 - P_2)\frac{1-\pi}{\pi}]}\right\}^2}{(P_1 - P_2)^2 (1 - \pi)}$$

Whittemore (1981) / Hsieh (1998)

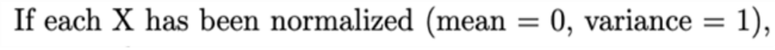


Let Y denote status.
Let Y = 1 if disease occurs at Y = 0.
Let X_1, ..., X_k denote covariates.

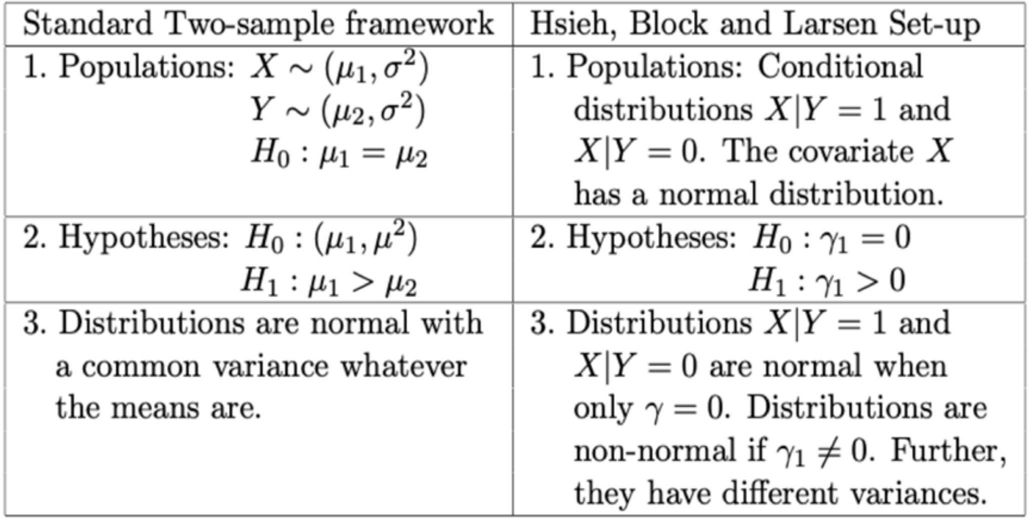
$P = P(X) = P(Y = 1|X_1, ..., X_k)$ is
$\log[P/(1\text{-}P)] = \theta_0 + \theta_1 X_1 + ... + \theta_k X_k$



H_0: $\theta = [0, \theta_2, ..., \theta_k]$
H_a: $\theta = [\theta^*, \theta_2, ..., \theta_k]$



If each X has been normalized (mean = 0, variance = 1),



Standard Two-sample framework	Hsieh, Block and Larsen Set-up
1. Populations: $X \sim (\mu_1, \sigma^2)$ $Y \sim (\mu_2, \sigma^2)$ $H_0 : \mu_1 = \mu_2$	1. Populations: Conditional distributions $X\|Y = 1$ and $X\|Y = 0$. The covariate X has a normal distribution.
2. Hypotheses: $H_0 : (\mu_1, \mu^2)$ $H_1 : \mu_1 > \mu_2$	2. Hypotheses: $H_0 : \gamma_1 = 0$ $H_1 : \gamma_1 > 0$
3. Distributions are normal with a common variance whatever the means are.	3. Distributions $X\|Y = 1$ and $X\|Y = 0$ are normal when only $\gamma = 0$. Distributions are non-normal if $\gamma_1 \neq 0$. Further, they have different variances.

	Whittemore/Hsieh	EPV
Pros	● No need for coefficient of determination of covariate ● Interim or group-sequential designs ● Smaller required sample size ● Direct and exact equation **Hsieh** ● No assumption of small response probability	● Sample size of 500 will yield reliable sample estimates ● Smaller sample size -> larger effect size therefore estimate an almost accurate effect size. ● Easy to calculate / concise equation
Cons	● Needed adjustment for sample size tables ● Not giving actual power close to nominal one as γ_0 increases **Hsieh** ● Very sensitive to choice of γ_0 ● Conditions for two-sample not exact to framework ● Failed when covariate X is bernoulli **Whittemore** ● Do not meet nominal levels of power for certain range of parameter values ● Can not ignore small response of probability condition	● Involves many parameters (difficult to estimate) ● Large sample size can introduce bias ● Too many variables may cause noise ● Complicated and iterative without explicit formula ● Validation equation was tested based on single dataset ● Simulation analysis was not conducted ● Does not take into account the power of the experiment.

Sample size Calculations for Logistic Regression, Binomial Equation

Binary case (i.e. independent variable is distributed as binomial)

- Minimum sample size required with a given β and α :

where α/2 = significance level

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{(p_1 - p_0)^2} \cdot \left[\frac{p_0(1-p_0)}{1-\pi} + \frac{p_1(1-p_1)}{\pi}\right]$$

- β = power
- π = portion of sample when x=1
- p_0 = P(y = 1 | x = 0)
- p_1 = P(y = 1 | x = 1)
- z_{χ} = corresponding Z table value to x

Sample size Calculations for Logistic Regression, Normal Equation (Contd.)

Normal case (i.e. the independent variable is distributed normally)

- The minimum sample size when comparing the null hypothesis Ho: $\beta_1 = 0$ with the alternative hypothesis Ha: $\beta_1 = b$ can be estimated by:

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{p_0(1 - p_0)b^2}$$

Where, - α / 2 = significance

- β = power

- b = alternative hypothesis estimate for β_1

- p_0 = P(y = 1 | x = μ_χ)

- z_χ = corresponding Z table value to x

How to use R packages, if there are any, to facilitate the calculation?

R Packages for Sample Calculation

- A very useful R packages in r used for sample size calculations is “pwr”. Here is the list of its following functions and required quantities:

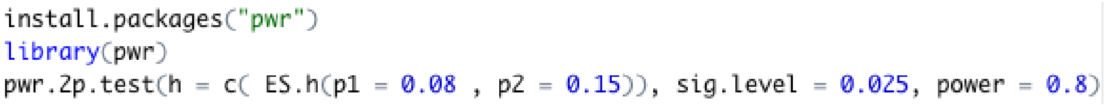
- For Logistic regression, we will only be able to use the following highlighted functions, as some, for example, use non-binary distributions parameters, where we are looking for binary models, i.e. from the binomial family, and even some normally distributed variables, as denoted for logistic regression in the form $Y = \beta_1 X + \beta_0$.

Function:	Power calculations for:
pwr.2p.test()	Two proportions (equal n)
pwr.2p2n.test()	Two proportions (unequal n)
pwr.anova.test()	Balanced one way Anova
pwr.chisq.test()	Chi-square test
pwr.f2.test()	Glm (generalized linear model)
pwr.p.test()	Proportion (one sample)
pwr.r.test()	Correlation
pwr.t.test()	T-test (one or two sample, paired)
pwr.t2n.test()	T-test (two samples with unequal n)

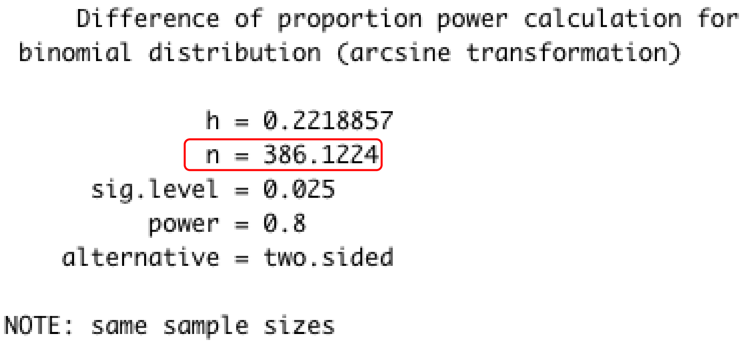
R Packages Cont.

- In this package, we will show an example of the following function pwr.2p.test. The function is used as follows pwr.2p.test (h = c(ES.h(p1 = , p2 =)), sig.level = , power = , alternative =).
- We will also show examples for functions:
- pwr.p.test. The function is used as follows pwr.p.test (h = c(ES.h(p1 = , p2 =)), sig.level = , power = , alternative =).
- pwr.chisq.test(w = , N = , df = , sig.level = , power =)
- pwr.r.test(r = , sig.level = , power =)

Example, 2 proportions, equal n



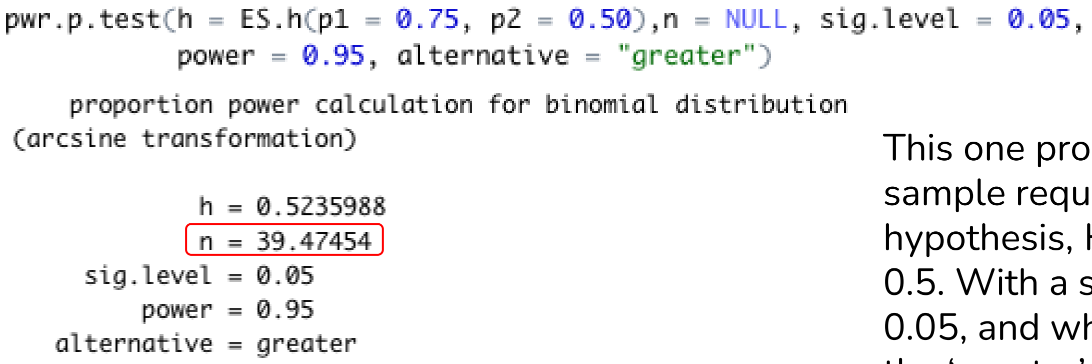
```
install.packages("pwr")
library(pwr)
pwr.2p.test(h = c( ES.h(p1 = 0.08 , p2 = 0.15)), sig.level = 0.025, power = 0.8)
```



```
     Difference of proportion power calculation for
 binomial distribution (arcsine transformation)

              h = 0.2218857
              n = 386.1224
      sig.level = 0.025
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

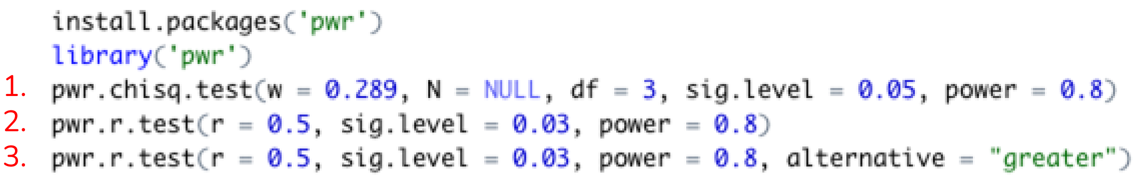
We would use this functions if we have two comparable proportions drawn from the same sample size. Given our power of 0.8, we determine a sample of n = 387(always rounding up), would suffice our 95% confidence interval. Function ES.h helps us calculate the effect size with the two given proportions.

Example, 1 proportion, equal n

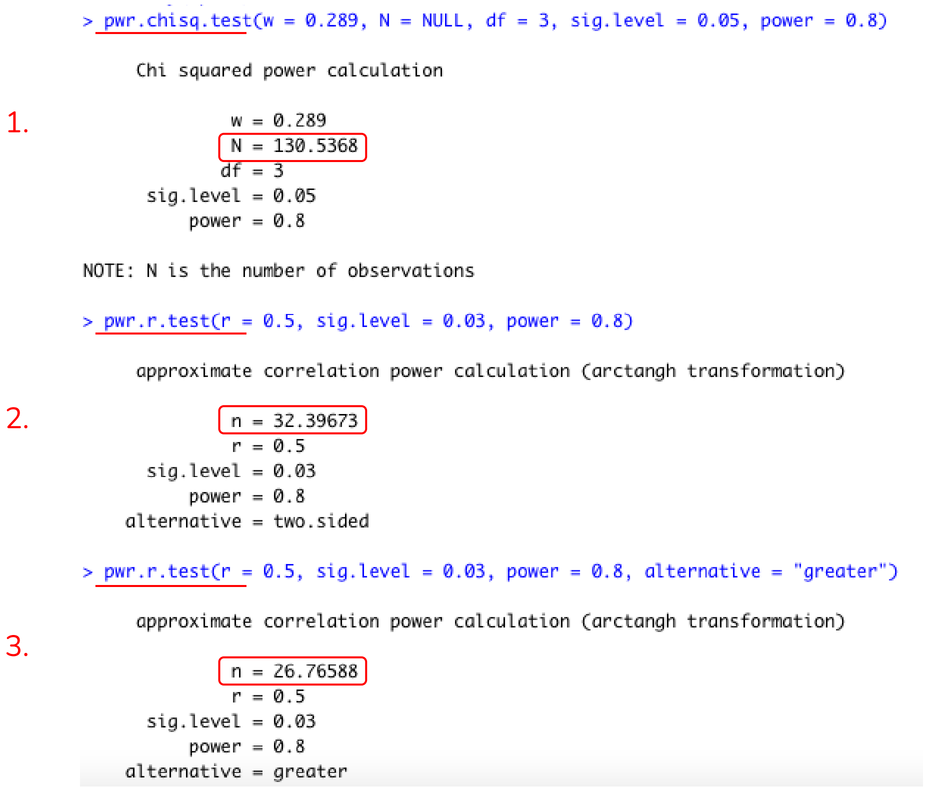


```
pwr.p.test(h = ES.h(p1 = 0.75, p2 = 0.50),n = NULL, sig.level = 0.05,
           power = 0.95, alternative = "greater")

     proportion power calculation for binomial distribution
 (arcsine transformation)

              h = 0.5235988
              n = 39.47454
      sig.level = 0.05
          power = 0.95
    alternative = greater
```

This one proportion test finds the sample required to test the hypothesis, Ho : p =0.75, Ha : p = 0.5. With a significance level of 0.05, and when ‘alternative’ set to the ‘greater’, meaning the greater half of the area under the curve, meaning significant level of 95%, we must at least sample the next n = 40 to test with 95% power.



```
    install.packages('pwr')
    library('pwr')
1.  pwr.chisq.test(w = 0.289, N = NULL, df = 3, sig.level = 0.05, power = 0.8)
2.  pwr.r.test(r = 0.5, sig.level = 0.03, power = 0.8)
3.  pwr.r.test(r = 0.5, sig.level = 0.03, power = 0.8, alternative = "greater")
```



```
> pwr.chisq.test(w = 0.289, N = NULL, df = 3, sig.level = 0.05, power = 0.8)

     Chi squared power calculation

1.            w = 0.289
              N = 130.5368
             df = 3
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations

> pwr.r.test(r = 0.5, sig.level = 0.03, power = 0.8)

     approximate correlation power calculation (arctangh transformation)

2.              n = 32.39673
                r = 0.5
        sig.level = 0.03
            power = 0.8
      alternative = two.sided

> pwr.r.test(r = 0.5, sig.level = 0.03, power = 0.8, alternative = "greater")

     approximate correlation power calculation (arctangh transformation)
3.
                n = 26.76588
                r = 0.5
        sig.level = 0.03
            power = 0.8
      alternative = greater
```

What are the key factors in sample size calculation?

Definition of a “key factor”?

- Let us say a key factor is important information to be aware of when calculating sample sizes
- Such important information includes potential problems, and their corresponding solutions

What is your intent?

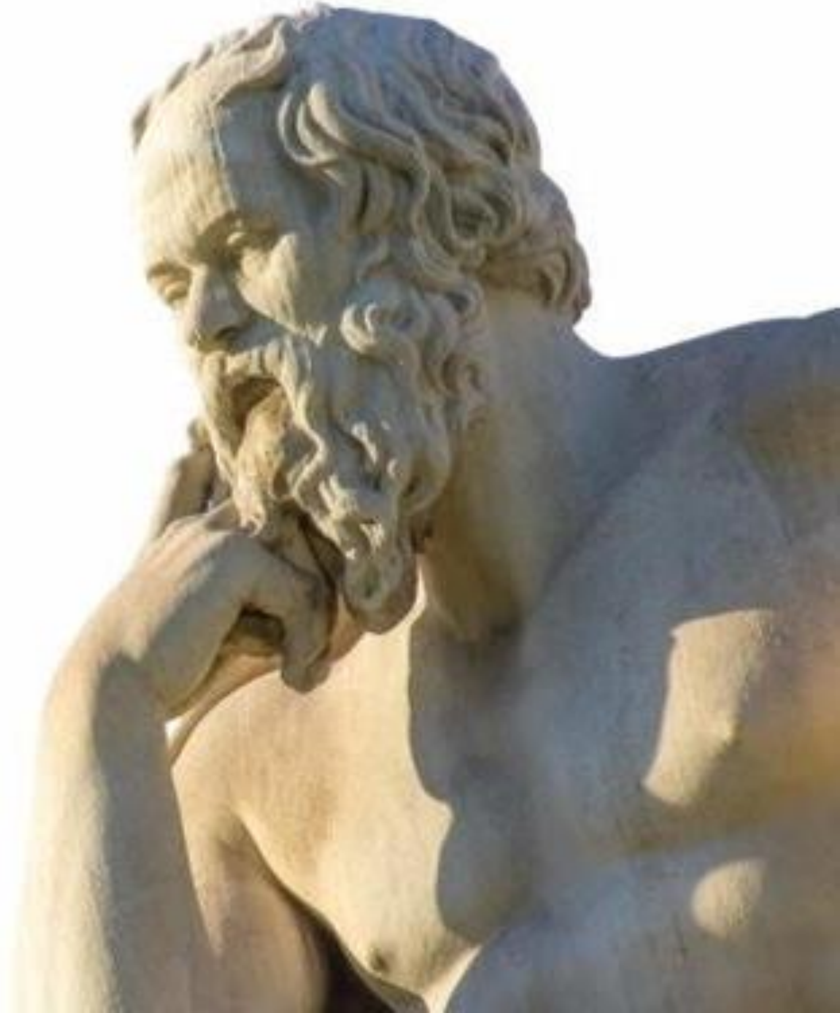
Industry
Academia
imgflip.com
JAKE-CLARK.TUMBLR

My intent

“**scientia**” - *(noun)* knowledge, and specifically for theoretical knowledge

“**ars**” - *(noun)* art – in the sense that applying knowledge (ie application) is an art in of itself

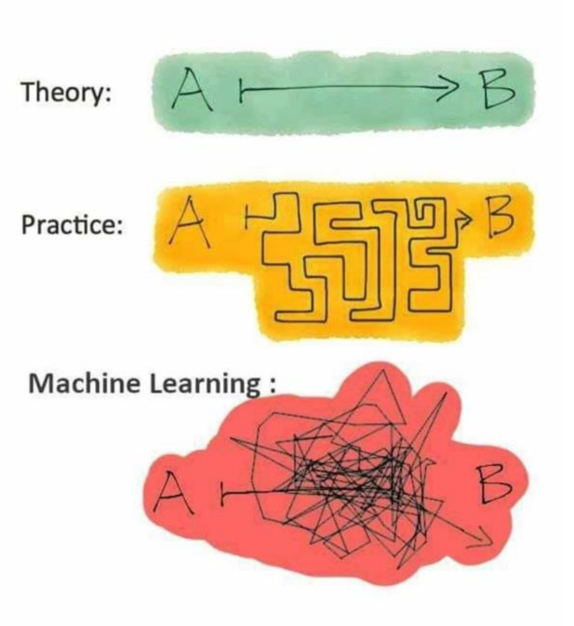
That is, knowing for the sake of doing, and not knowing, in order to, sorry, make money.



Theory vs practice

“In theory, theory and practice are the same 100% of the time. In practice, theory and practice are the same never of the time.”

– Best piece of advice I heard when I was in engineering



What
Do I Do
Now

StackExchange

Quick-R by datacamp

Wikipedia
The Free Encyclopedia

stack overflow

RDocumentation

Where I head when I need advice I actually will use: Why read dense academic articles when I can get the run-down someone else has already prepared?

Possible uses of logistic modelling (outside biostatistics)

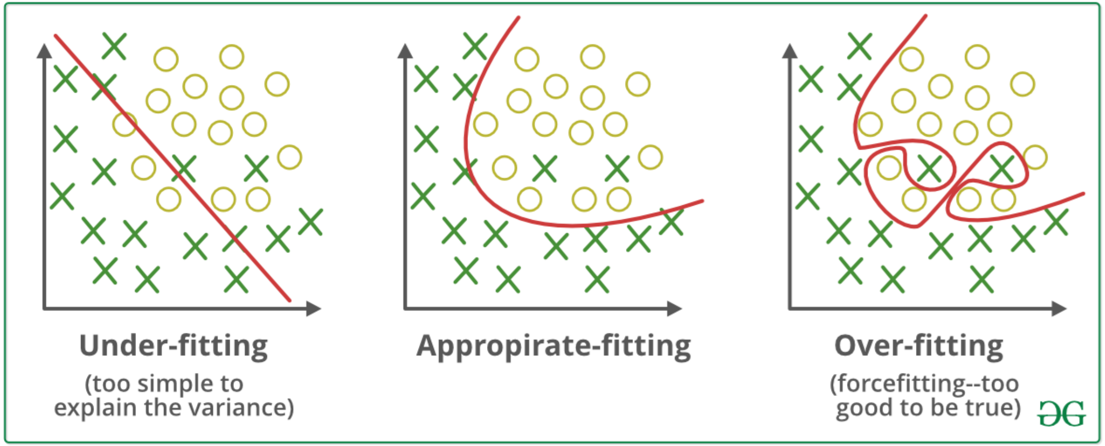
“ For example, whether a voter will vote for a Democrat or Republican can be determined through the interpretation of logistic modeling, which is based on demographic parameters such as gender, age and the state of residence of the voter..”

“In business, logistic regression analysis is suited to data mining applications which are used in business analytics. For example, logistic regression modeling can be used to predict customer retention, such as a yes/no/maybe scenario indicating, whether a customer will re-visit/not visit again/may visit again based on the marketing stimuli.”

What I would do to

- Determine if a voter will vote for Democrats or Republicans in a US election
    1. Look at previous study designs (for inspiration)
    2. Choose variables of interest: What do I care about? Sex? Ethnicity? Income? State? Interactions between 2+ predictors?
    3. Sample as many voters as possible
    4. Enter the code into a .csv file
    5. Run the glm() function in R to perform a logistic regression, but only on 75-80% of the entries, to which I give the name “training data”
    6. Procure the model
    7. Look at ROC curves given in R
    8. Examine possible issues (see next few slides)
    9. Run it on the remaining 20-25% of the entries, to which I give the name “testing data”
    10. Determine how well the model classified the “testing data” – Should I care about EPV if prediction appears to work well?

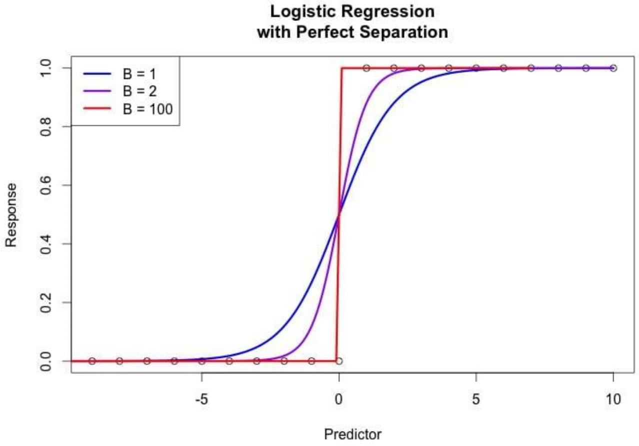
Underfitting vs. overfitting

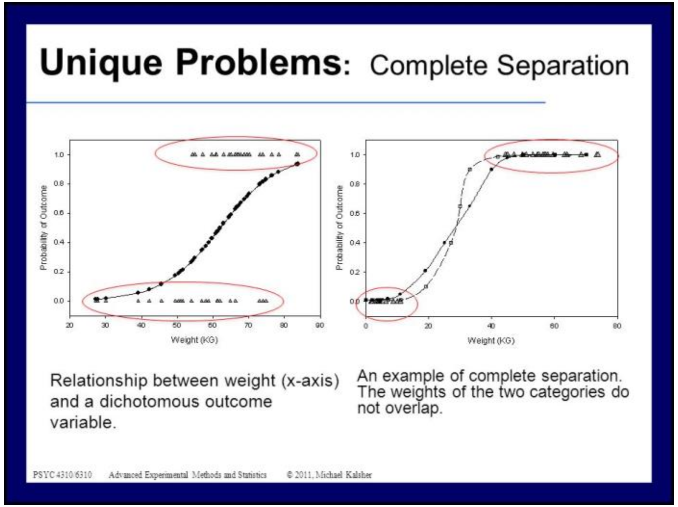

Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)
ƏG

- Underfitting occurs as EPV (variance) is too large
  - Possible solutions include increasing model complexity
- Overfitting (no randomness) occurs as EPV approaches 1

Separation

- 1's gathered on one extreme of an independent variable (or some combination of them), and all of the 0's at the other extreme.
- Although this would seem like a good situation, because it would make perfect prediction easy, it ruins parameter estimation
- Possible solutions:
  - First - ask yourself, is the separation a byproduct of your sample, or true in the population?
  - Penalized regression
  - Exclude cases
  - Do nothing – an analysis could also reveal something





Unique Problems: Complete Separation

Probability of Outcome
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
80
90
Weight (KG)
Probability of Outcome
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Weight (KG)

Relationship between weight (x-axis) and a dichotomous outcome variable.

An example of complete separation. The weights of the two categories do not overlap.

PSYC 4310/6310 Advanced Experimental Methods and Statistics © 2011, Michael Kalsher

Conclusion

Remarks

Sample size calculation can be incredibly cumbersome, but one can employ tricks to save time, energy – and brain power, culminating in a beautiful result.

Sources

A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. Peduzzi et al. (1996)

https://rdrr.io/cran/powerMediation/man/SSizeLogisticBin.html

https://www.statmethods.net/stats/power.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5394463/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6422534/

https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

https://rdrr.io/cran/powerMediation/man/SSizeLogisticBin.html

http://www.statpower.net/Content/312/Handout/Hsieh%281989%29.pdf

https://www.researchgate.net/publication/13586507_A_Simple_Method_of_Sample_Size_Calculation_for_Linear_and_Logistic_Regression

https://link-springer-com.ezproxy.lib.ucalgary.ca/content/pdf/10.1007/s13571-010-0004-6.pdf

Sample size for logistic regression? - Cross Validated (stackexchange.com)

r - How to deal with perfect separation in logistic regression? - Cross Validated (stackexchange.com)

https://bookdown.org/pdr_higgins/rmrwr/sample-size-calculations-with-pwr.html

https://www.real-statistics.com/logistic-regression/logistic-regression-sample-size/logistic-regression-sample-size-binary/

https://www.real-statistics.com/logistic-regression/logistic-regression-sample-size/