Analyses in Reproducible Research:

An Example with Logistic Regression

Ryan T. Pohlig

Presentation Overview

- Logistic Regression
  - Overview
- Brief Intro to SPSS
  - Using Syntax in SPSS
- Example Research
  - Quick Overview
  - Talk about the data
- Running the analysis in SPSS
  - Checking assumptions
  - Interpreting results

- DV and Outcome are going to be used interchangeably
- IV and Predictor will be used interchangeably

Logistic Regression

Crash Course

Logistic regression

- Conceptually it is no different than the multiple egression you (may) be familiar with
- Research Questions are even the same
  - Does an IV predict a DV?
  - Does a set of IVs predict a DV?
  - Does an IV predict a DV, after adjusting for covariates?
- Differences:
  - When an DV is categorical, cannot use multiple regression
  - Interpretation of coefficients are different

Categorical DVs

- Different Methods for analysis given a categorical DV
  - Discriminant Function Analysis
    - Part of MANOVA family (IV’s need to be interval or ratio)
  - Loglinear Modeling (Multi-way Frequency Analysis)
    - Generalization of two-way chi-square test (or more dimensions than Mantel-Haenszel test)
  - Generalized Linear Modeling
    - Generalization of the more familiar General Linear Modeling (ANOVA/Regression)
    - Specify a link function & a distribution from the exponential family
    - More flexible, can accommodate many types of outcomes
      - Probit Regression or Logistic Regression
      - Poisson Regression

LR : Regression

- Predictors & IVs function the same as in regression
  - Can have continuous predictors
  - Categorical ones must be dummy or contrast coded
  - Can be sequential with tests for multiple blocks of predictors
  - Can have Automated selection procedures
  - Can test moderation using interactions

- Technically LR is regression looking at a non-linear relationship between predictors and outcomes
- One can linearize arelationship by
  - Applying a transformation to the outcome
  - Applying a transformation to the predictors

LR : MR part 2

- MR formula in matric notation
  - $y = \beta X + \varepsilon$
    - y n by 1 vector of observed DV
    - β p by 1 vector of regression coefficients
    - X n by p matrix of observed IVs
    - ε n by 1 vector of error terms
- Applying transformation to the outcome
  - $g(y) = \beta X + \varepsilon$
- Applying Transformation to predictors & errors
  - $y = f(\beta X + \varepsilon)$

- In LR, the logit function is used
  - $g(y) = logit = \log\left(\frac{\pi}{1-\pi}\right)$
- Where π is the probability y=1 given your predictors
  - Formally: $\pi = P(y = 1|X = x)$
- Where 1-π is the probability y=0 given your predictors
  - Formally: $1 - \pi = P(y = 0|X = x)$

LR Model

- In logit form
  - $g(y) = \log\left(\frac{\pi}{1-\pi}\right) = X\beta + \varepsilon$
- Solve for π
  - $\pi = \frac{e^{x\beta+\varepsilon}}{1+e^{x\beta+\varepsilon}}$
- For predicted model
  - $g(\hat{y}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = X\hat{\beta}$
- Solve for π
  - $\hat{\pi} = \frac{e^{x\hat{\beta}}}{1+e^{x\hat{\beta}}}$
- In order to solve for $\hat{\beta}$, need to use maximum likelihood
  - Iterative process that maximizes the sample values to represent the population parameters
- Likelihood – probability density function
  - $\ell(\hat{\beta}) = \prod_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-\hat{y}_i)}$
- Log-likelihood function
  - $L(\hat{\beta}) = \log\left(\ell(\hat{\beta})\right) = \sum_{i=1}^{n}[y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)]$

LR

- Using this you can get a global likelihood ratio statistic for testing the entire model
  - $L(\hat{\beta})$ can be used in the log-likelihood ratio test
- Also individual predictors can be tested using a Wald chi-square test
  - $W_j = \left[\hat{\beta}_j / SE(\hat{\beta}_j)\right]^2$, where j refers to the predictor of interest
- LR Assumptions:
  - Independence
  - No multicollinearity
  - No outliers/influential cases
  - Linearity of continuous IV's with logit

Example

Smoke

		No	Yes	*Total*
Weight	**Norm**	223	114	337
	Low	70	81	151
	Total	293	195	488

- DV = Birth Weight
  - Low or Not
- IV = Mother Smoking during pregnancy
  - Yes or No

- Odds
  - Prob. of Low Weight for non-smokers?
    - $\frac{70/293}{223/293} = \frac{.239}{.761} = .314$
    - 31.4% chance of low birth weight for non-smokers
  - For smokers?
    - $\frac{81/195}{114/195} = \frac{.415}{.585} = .711$
    - 71.1% chance of low birth weight for smokers
- Odds ratio
  - $\widehat{OR} = \frac{.314}{.711} = 2.26$
  - 2.3 times more likely to have a low birth weight, if the mother smoked during pregnancy

Logistic Regression Results

- Logistic Regression results
  - Given you coded smoking as 0 for no, 1 for yes
  - $\widehat{Logit} = -1.159 + \ .817(smoke)$
- The betas are not interpreted directly
- β_j is amount of change in logit for each one unit change in X_j
- The null-hypothesis is essentially testing if the Odds Ratio differs than 1

- In order to make the parameter estimates interpretable, we need to exponentiate them
  - Intercept is probability of low birth weight given no smoking
    - $e^{\widehat{\beta}_0} = e^{-1.159} = .314$
  - Full equation is probability of low birth weight given smoking
    - $e^{\widehat{\beta}_0} = e^{-1.159+.817} = .710$
  - Slope is Odds Ratio, the impact of smoking on birth weight
    - $e^{\widehat{\beta}_1} = e^{.817} = 2.264$

LR- is your model good?

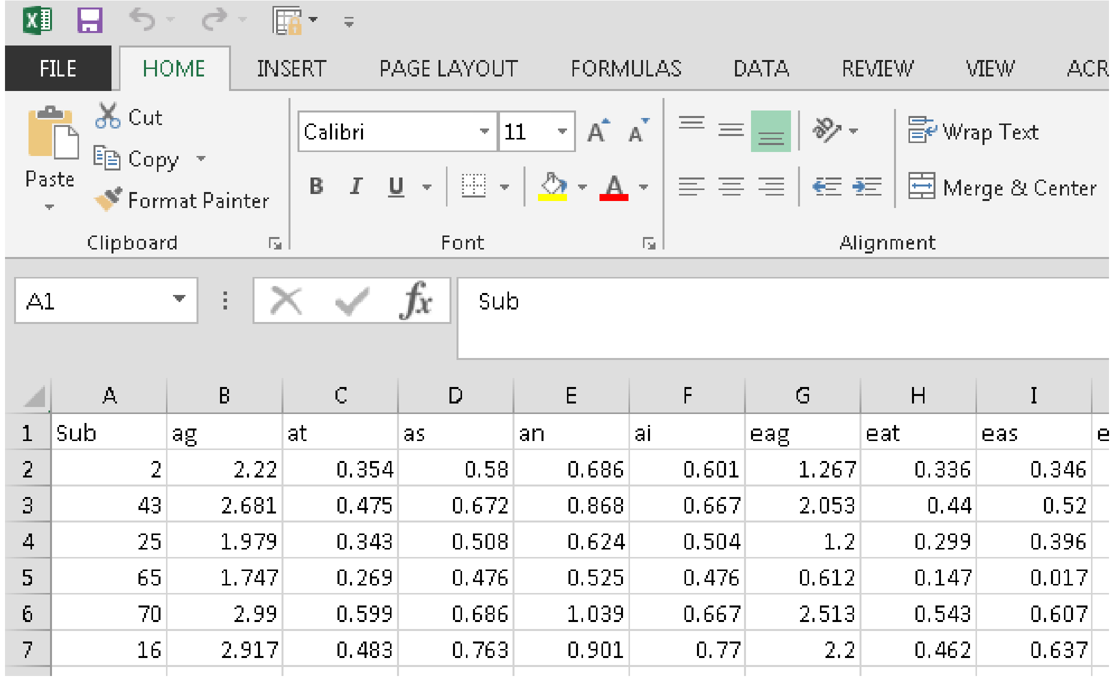
- There are different estimates for in R^2 LR (yes more than one way exists)
- Psudeo R^2 are used
  - http://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/
- Cox & Snell
  - Looks at improvement of predictors relative to a null model (intercept only)
- Negelkerke
  - Adjustment to Cox & Snell so that the max value can be 1
- Can look at Information Criteria to determine fit
  - AIC, BIC, etc.
  - Smaller of these indicates best fit
- Can calculate predicted probabilities for individuals' or use regression formula for a specific set of X values
- Hosmer-Lemeshow is another method for showing goodness of fit
- Many Use an ROC curve for significant predictors
  - Can look at Area Under the Curve
  - Youden's Index or Kullback-Leibler distance
  - Maximum value of Youden's index is the cut point for classification that minimizes both false negatives & false positives
    - Is Sensitivity + (Specificity-1)

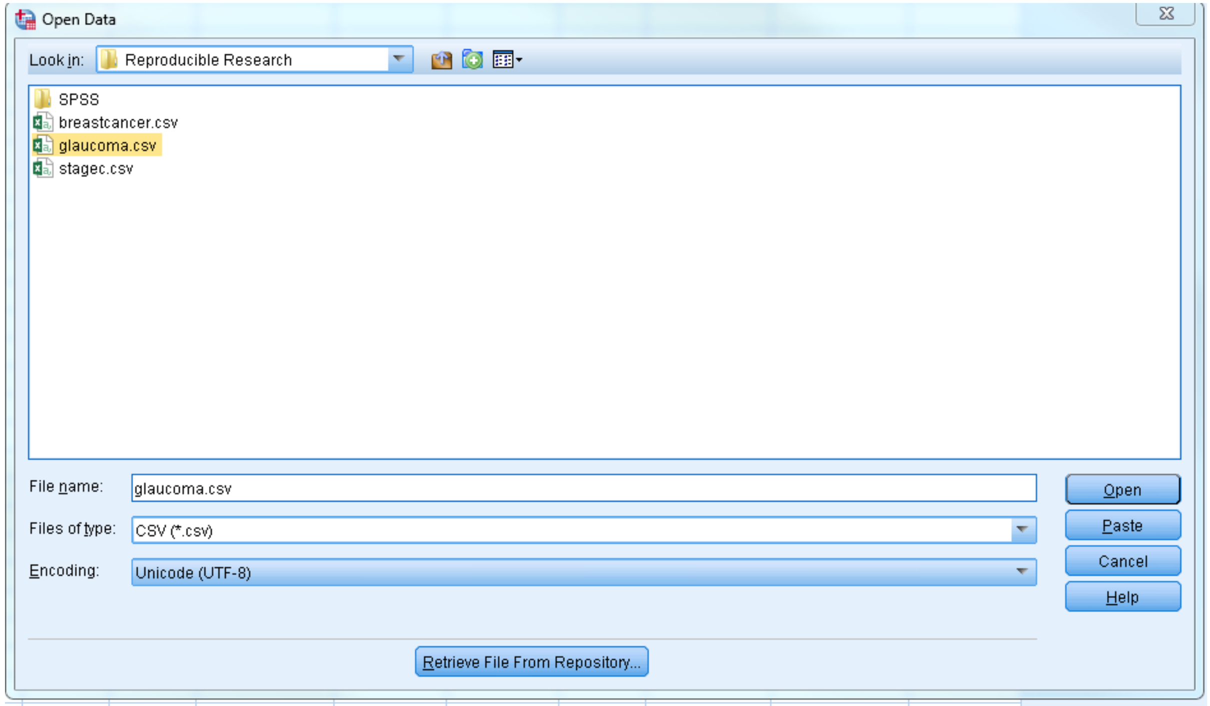
SPSS

Brief Intro

Check Data file

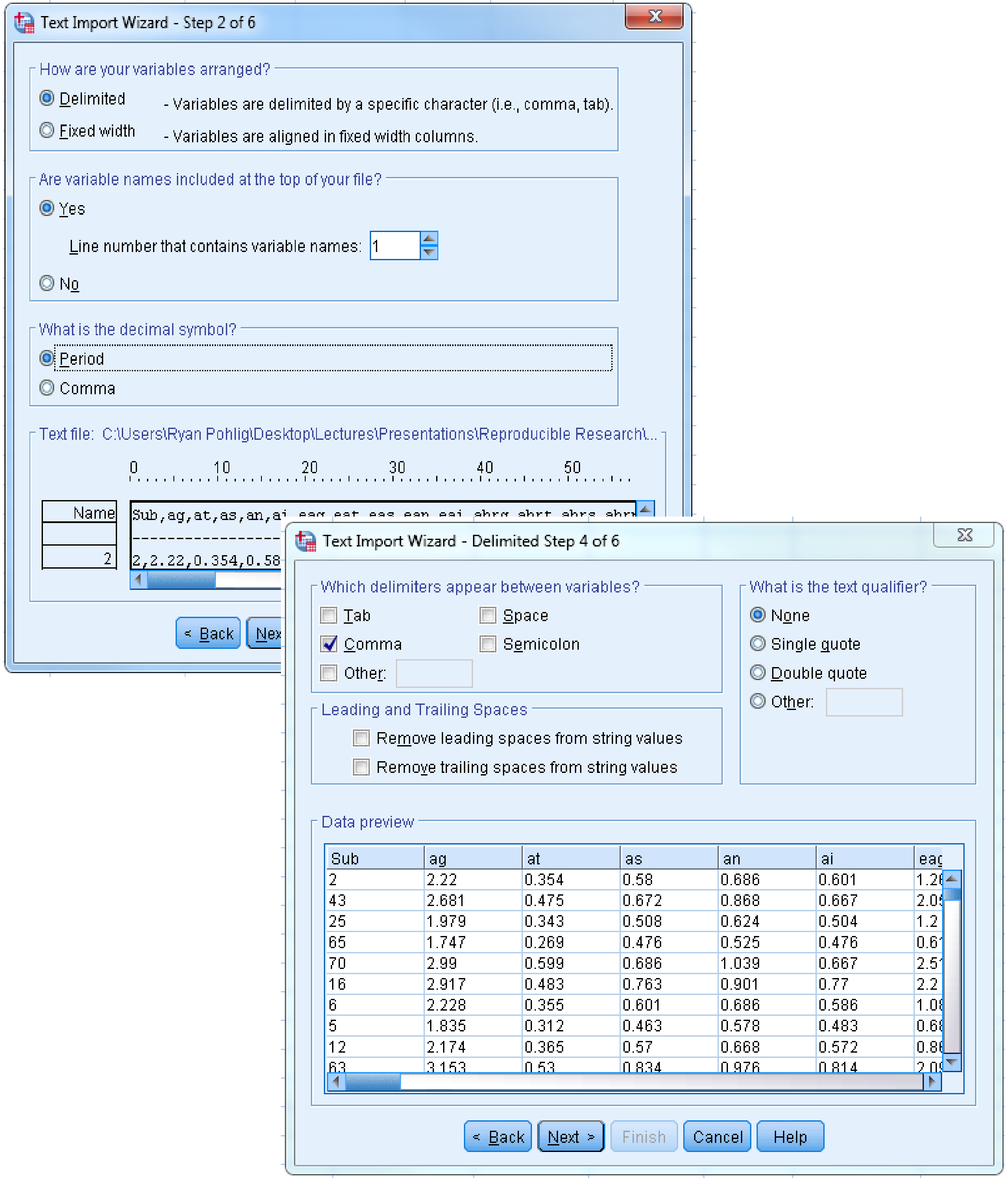
- Can Open .csv files with excel pretty easily
- Notice we have a header row with variable names
- Open SPSS
  - Choose File -> Open -> Data
  - Navigate to the correct folder
  - Make sure to choose .csv as file type
  - Hit 'open'





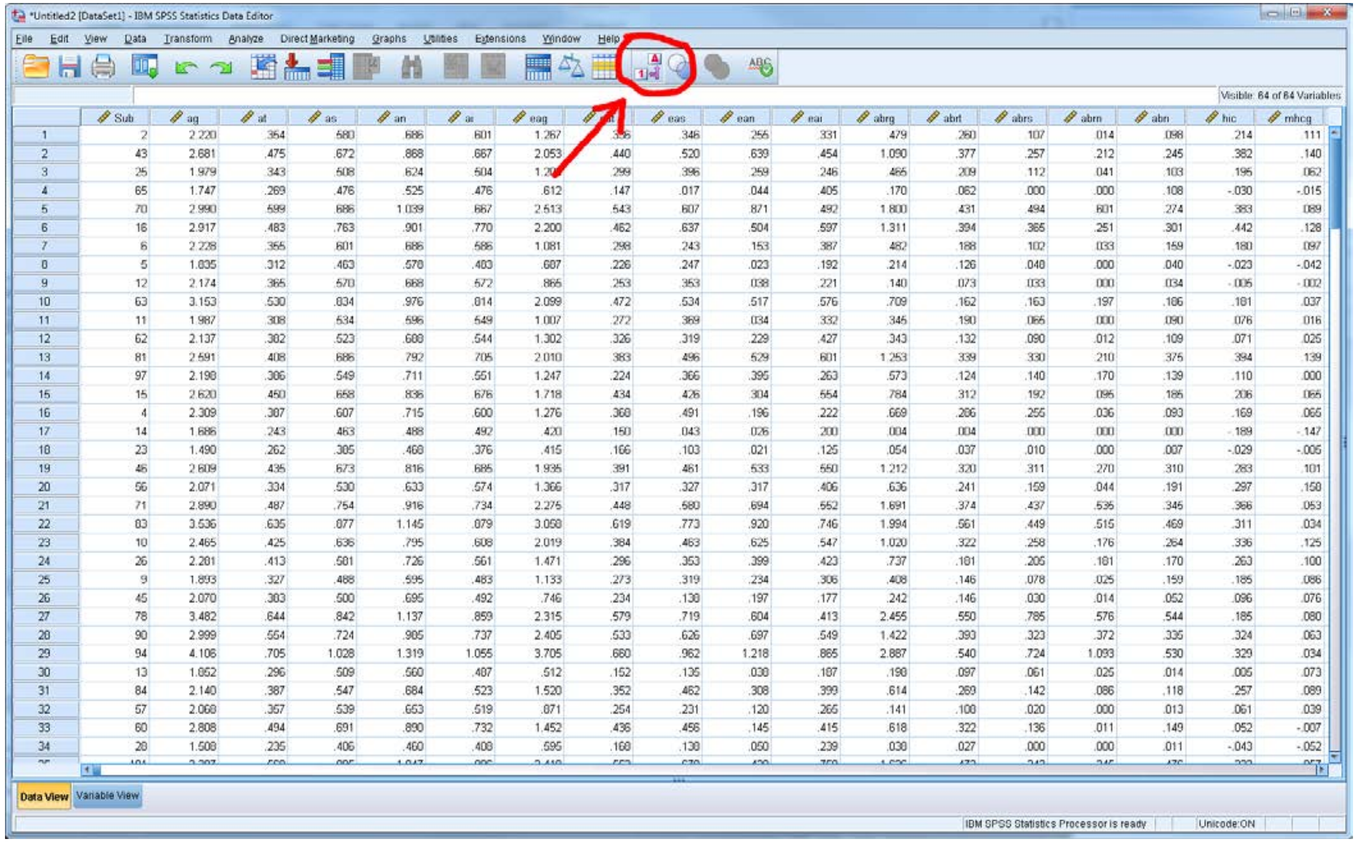
Importing Data

- Point & Click
  - SPSS does a good job at naively having the correct settings
  - Choose Delimited
    - With a CSV file, a comma is the delimiter
  - Make sure the top row is the variable names
  - Hit Next, Hit Next for Step 3
  - Make sure to choose Comma in step 4
  - Hit next all the way through to go right to the data
    - Alternatively choose to ‘paste syntax’
  - This opens a syntax file that can be used to skip the import data wizard in the future (Please save the syntax file now)



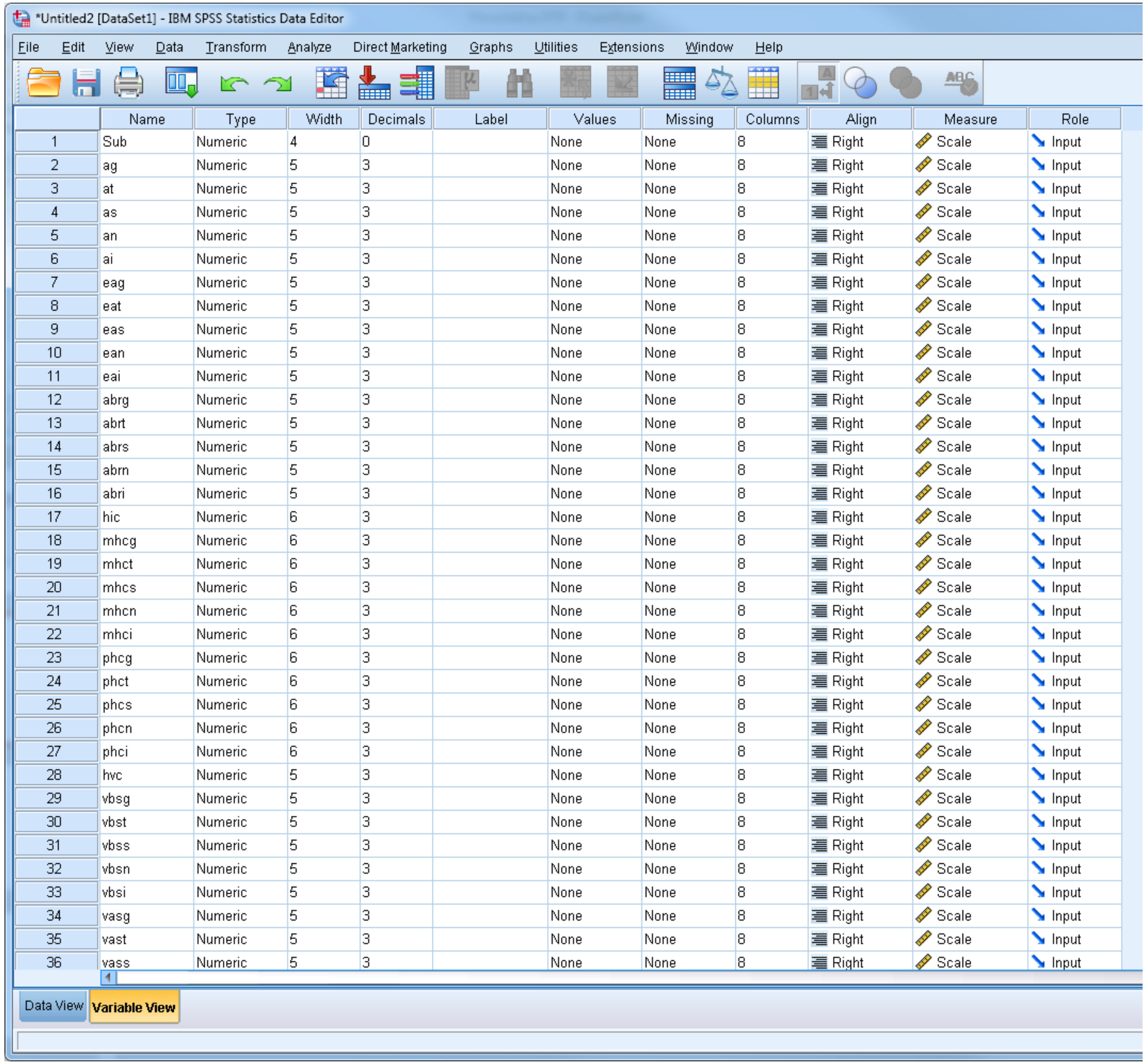
SPSS has 3 main Windows

- **Data Window** has 2 tabs
- The Data View
  - Looks like an excel file of your data
  - Contains your actual raw data
  - Notice the button at the top that has arrows pointing to an ‘A’ and a ‘1’
  - This button is used for showing variable labels
  - Do you want to see what the raw numeric values (0 or 1)
  - Or what those raw values stand for (Male or Female)



SPSS' Windows

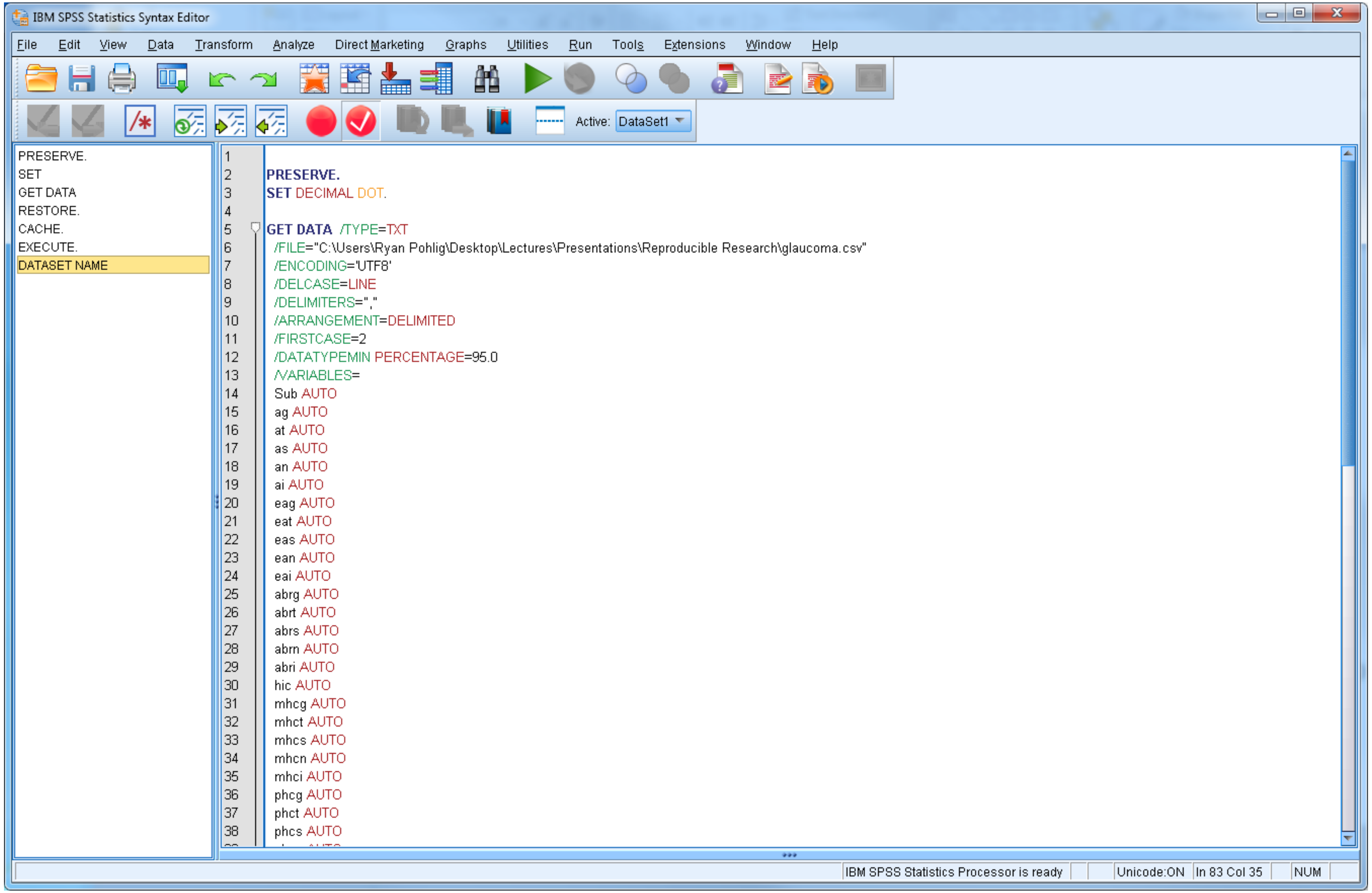
- **Data Window** has 2 tabs
- The Variable View
    - Contains information for each variable
    - Here you can label Variables
    - Label values (name what the indicators mean)
    - Alternatively You can do this in syntax
- You will notice a variable at the end called 'Class' that is not numeric



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Sub	Numeric	4	0		None	None	8	Right	Scale	Input
2	ag	Numeric	5	3		None	None	8	Right	Scale	Input
3	at	Numeric	5	3		None	None	8	Right	Scale	Input
4	as	Numeric	5	3		None	None	8	Right	Scale	Input
5	an	Numeric	5	3		None	None	8	Right	Scale	Input
6	ai	Numeric	5	3		None	None	8	Right	Scale	Input
7	eag	Numeric	5	3		None	None	8	Right	Scale	Input
8	eat	Numeric	5	3		None	None	8	Right	Scale	Input
9	eas	Numeric	5	3		None	None	8	Right	Scale	Input
10	ean	Numeric	5	3		None	None	8	Right	Scale	Input
11	eai	Numeric	5	3		None	None	8	Right	Scale	Input
12	abrg	Numeric	5	3		None	None	8	Right	Scale	Input
13	abrt	Numeric	5	3		None	None	8	Right	Scale	Input
14	abrs	Numeric	5	3		None	None	8	Right	Scale	Input
15	abrn	Numeric	5	3		None	None	8	Right	Scale	Input
16	abri	Numeric	5	3		None	None	8	Right	Scale	Input
17	hic	Numeric	6	3		None	None	8	Right	Scale	Input
18	mhcg	Numeric	6	3		None	None	8	Right	Scale	Input
19	mhct	Numeric	6	3		None	None	8	Right	Scale	Input
20	mhcs	Numeric	6	3		None	None	8	Right	Scale	Input
21	mhcn	Numeric	6	3		None	None	8	Right	Scale	Input
22	mhci	Numeric	6	3		None	None	8	Right	Scale	Input
23	phcg	Numeric	6	3		None	None	8	Right	Scale	Input
24	phct	Numeric	6	3		None	None	8	Right	Scale	Input
25	phcs	Numeric	6	3		None	None	8	Right	Scale	Input
26	phcn	Numeric	6	3		None	None	8	Right	Scale	Input
27	phci	Numeric	6	3		None	None	8	Right	Scale	Input
28	hvc	Numeric	5	3		None	None	8	Right	Scale	Input
29	vbsg	Numeric	5	3		None	None	8	Right	Scale	Input
30	vbst	Numeric	5	3		None	None	8	Right	Scale	Input
31	vbss	Numeric	5	3		None	None	8	Right	Scale	Input
32	vbsn	Numeric	5	3		None	None	8	Right	Scale	Input
33	vbsi	Numeric	5	3		None	None	8	Right	Scale	Input
34	vasg	Numeric	5	3		None	None	8	Right	Scale	Input
35	vast	Numeric	5	3		None	None	8	Right	Scale	Input
36	vass	Numeric	5	3		None	None	8	Right	Scale	Input

SPSS' Windows

- Output windows displays the results/output and contains any error messages
- Syntax window is the part you can directly manipulate code
- Comments will appear in gray on the left while commands appear in black



SPSS Syntax

- The majority of functions can be done via point + click/menu system
- In order for people to exactly replicate what you did ALWAYS paste it into syntax
- You can leave comments in syntax using '*'
- Syntax is often easier to manipulate when things need to be repeated
- SPSS needs each line of code to end with a '.' and then to run it you need to add 'execute.' after the commands
- Just highlight and hit run to do it
- SPSS will attempt to help you by showing valid commands in navy and errors in red
- Often times functions and analyses in SPSS are restricted by the 'type'
  - I only use numeric & string

SPSS

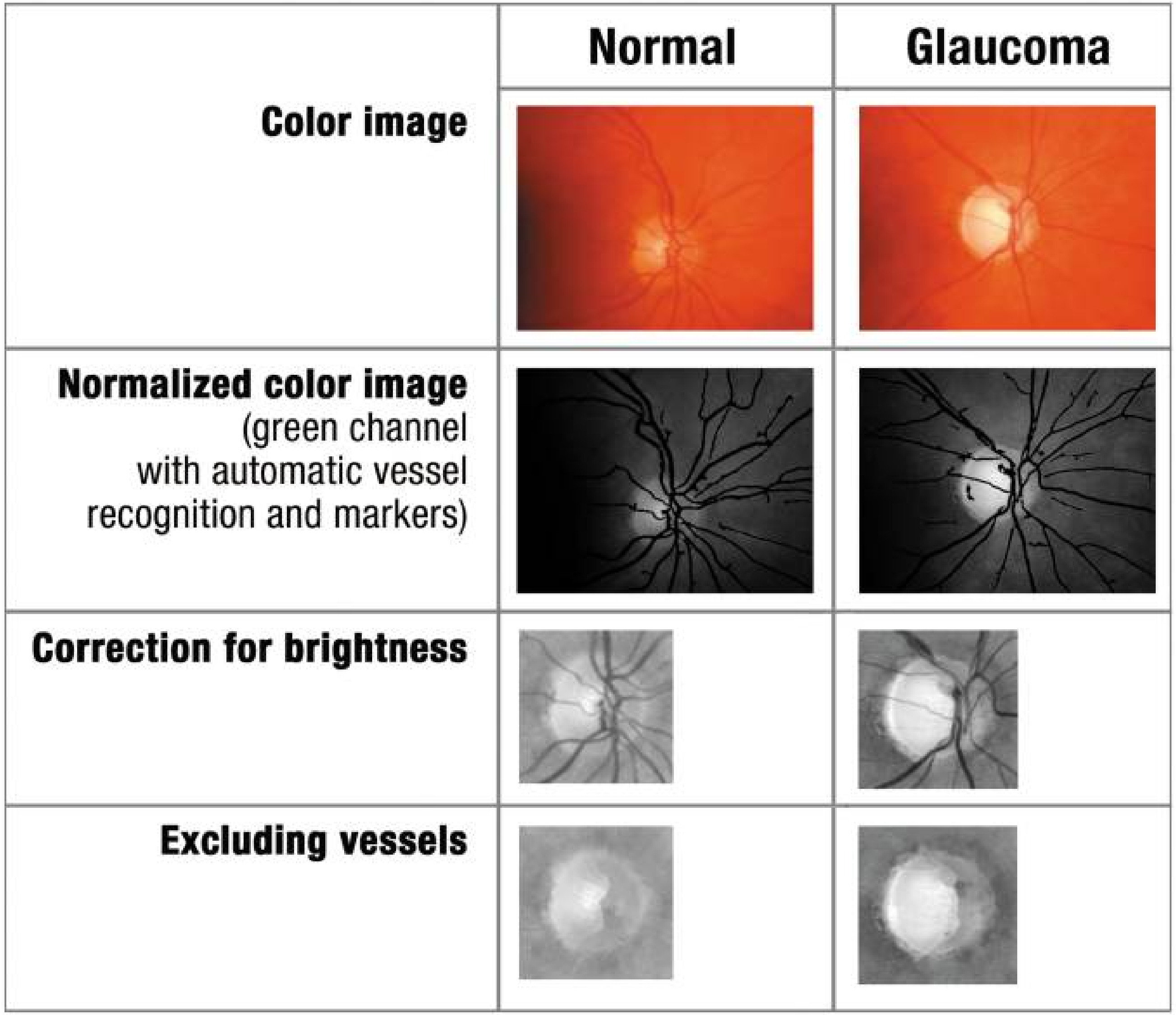
Brief Background & Example Analysis

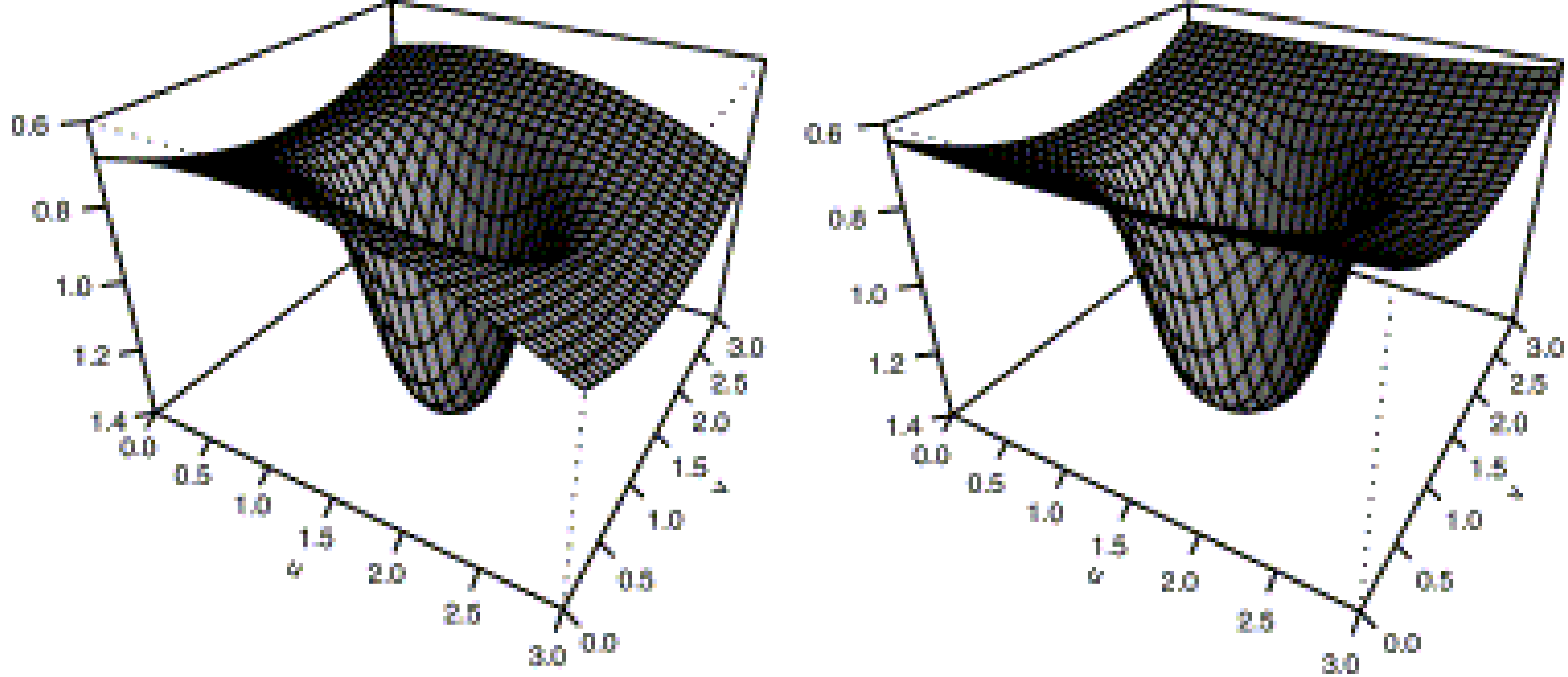
Analysis

- RQ: Predicting who might have glaucoma given measures of the optic nerve from a confocal laser scan (Heidelberg Retina Tomograph)
  - Data contains a variable with clinical diagnosis of glaucoma or normal
  - Area & Volume measures for parts of the optic nerve head measured in four locations (temporal, superior, nasal, and inferior)
  - Additionally have global measures which are sum of all four sectors
- Data & area image taken from:
  - Hothorn & Berthold Lausen (2003). Double-Bagging: Combining classifiers by bootstrap aggregation. Pattern Recognition, 36(6), 1303–1309
- Background & nerve image taken from:
  - Michelson, G., Hornegger, J., Wärntges, S., & Lausen, B. (2008). The Papilla as Screening Parameter for Early Diagnosis of Glaucoma. *Deutsches Ärzteblatt International*, *105*(34-35), 583–589

Brief Background

- Roughly 67 million people suffer from glaucoma (third most common cause of blindness)
- Early diagnosis of glaucoma, is essential because by the time the patient notices functional impairment, the damage is irreversible
- Early treatment can decrease the rate of blindness 20 years later by about 50%
- As a result of the pressure, an excavation (cupping) of the optic disc may occur, along with diminution of the visual field





- Surface of the optic nerve head model for normal (left) and glaucomatous (right)
- Can sort of see that papillary excavation for the glaucomatous eye is a bit different

Analysis

- Can we predict which eye has glaucoma based on area & volume measures from the scan?
- Specifically can these three measures predict Glaucoma?
  - Effective Area in Nasal sector (EAN)
  - Peak Height Contour in Temporal (PHCT)
  - Overall Mean Radius (MR)
- Before we start, we should check to see if anyone has started working on the data set
  - I see Melissa started examining the data in R, we can open her knitted file and see what she started doing!
    - Looking through her notes, we find “Note: Per PI, subject 91 non-compliant”

In SPSS

- We will use Syntax in SPSS for data manipulation
  - Creating a new variable called status that is 0 or 1 for being normal or having glaucoma
    - We can give it a variable label
    - Can apply value labels to the coding scheme of 0 or 1