The Impact of Socioeconomic Status and Education on the Incidence of Alzheimer's Disease

Data 621 Winter 2022 Course Project

Kimberley Chiu, Marc McCoy, Mark Ly

2022-04-02

Abstract

Background: Noise, as the term itself suggests, is most often seen a nuisance to ecological insight, a inconvenient reality that must be acknowledged, a haystack that must be stripped away to reveal the processes of interest underneath. Yet despite this well-earned reputation, noise is often interesting in its own right: noise can induce novel phenomena that could not be understood from some underlying determinstic model alone. **Methods:** Nor is all noise the same, and close examination of differences in frequency, color or magnitude can reveal insights that would otherwise be inaccessible. **Results:** Yet with each aspect of stochasticity leading to some new or unexpected behavior, the time is right to move beyond the familiar refrain of "everything is important" (Bjørnstad & Grenfell 2001). Stochastic phenomena can suggest new ways of inferring process from pattern, and thus spark more dialog between theory and empirical perspectives that best advances the field as a whole.
Conclusion: I highlight a few compelling examples, while observing that the study of stochastic phenomena are only beginning to make this translation into empirical inference. There are rich opportunities at this interface in the years ahead. Abstract word count: 350 Manuscript word count: 1500 Figures: 7 figures Tables: 7 tables References:

Contents

1 Introduction: Noise the nuisance

To many, stochasticity, or more simply, noise, is just that – something which obscures patterns we are trying to infer (Knape & de Valpine 2011); and an ever richer batteries of statistical methods are developed largely in an attempt to strip away this undesirable randomness to reveal the patterns beneath (Coulson 2001). Over the past several decades, literature in stochasticity has transitioned from thinking of stochasticity in such terms; where noise is a nuisance that obscures the deterministic skeleton of the underlying mechanisms, to the recognition that stochasticity can itself be a mechanism for driving many interesting phenomena (Coulson et al. 2004). Yet this transition from noise the nuisance to noise the creator of ecological phenomena has had, with a few notable exceptions, relatively little impact in broader thinking about stochasticity. One of the most provocative of those exceptions has turned the classical notion of noise the nuisance on its head: recognizing that noise driven phenomena can become a tool to reveal underlying processes: to become noise the informer. Here I argue that this third shift in perspective offers an opportunity to better bridge the divide between respective primarily theoretical and primarily empirical communities by seeing noise not as mathematical curiosity or statistical bugbear, but as a source for new opportunities for inference.

2 Methods

2.1 Study Design

In arguing for this shift, it essential to recognize this is a call for a bigger tent, not for the rejection of previous paradigms. What I will characterize as 'noise the nuisance' reflects a predominately statistical approach, in which noise, almost by definition, represents all the processes we are not interested in that create additional variation which might obscure the pattern of interest. By contrast, an extensive literature has long explored how noise itself can create patterns and explain processes from population cycling to coexistence. These broad categories should be seen as a spectrum and not be mistaken for either a sharp dichotomy nor a reference to a strictly empirical-theoretical divide. Each paradigm expands upon rather than rejects the previous notion of noise: the recognition that noise can create novel phenomena does not mean that noise cannot also obscure the signal of some process of interest. Likewise, seeking to use noise as a novel source of information about underlying processes will be informed by both previous paradigms, as our discussion will illustrate.

2.2 Study population

The dataset contains 9 demographic, clinical and derived anatomic values for 150 patients. Originally, the dataset was from a longitudinal study however, we plan to use it for this for a logistic regression. To satisfy the independence assumption, we will only use information from the first visit in our analysis. Covariates include sex (M/F), age (60 - 96), years of education (6 - 23), social economic status (SES; ranked 1-5), mini-mental stat examination (MMSE), clinical dementia rating (CDR), estimated total intracranial volume (eTIV), atlas scaling factor (ASF), and normalized whole brain volume (nWBV). Each unique patient was grouped into one of three categories, Nondemented, Demented, and Converted. Subjects that did not have a SES value recorded (8) were excluded from this study. Outliers were identified by checking the distribution. Figure(#) shows the details of our screening and inclusion process.

2.3 Study Outcomes

2.3.1 Primary Outcome

To emphasize the underlying trend in the changing roles in which we see and understand noisy processes, I will also restrict my focus to relatively simple models primarily from population ecology context. Simplicity not only makes examples (in equations and in code) more tractable but also allows us to focus on aspects that are germane to many contexts rather than unique to particular complexities (Bartlett 1960; Levins 41 1966).

2.3.2 Secondary Outcome

2.4 Data collection and data source

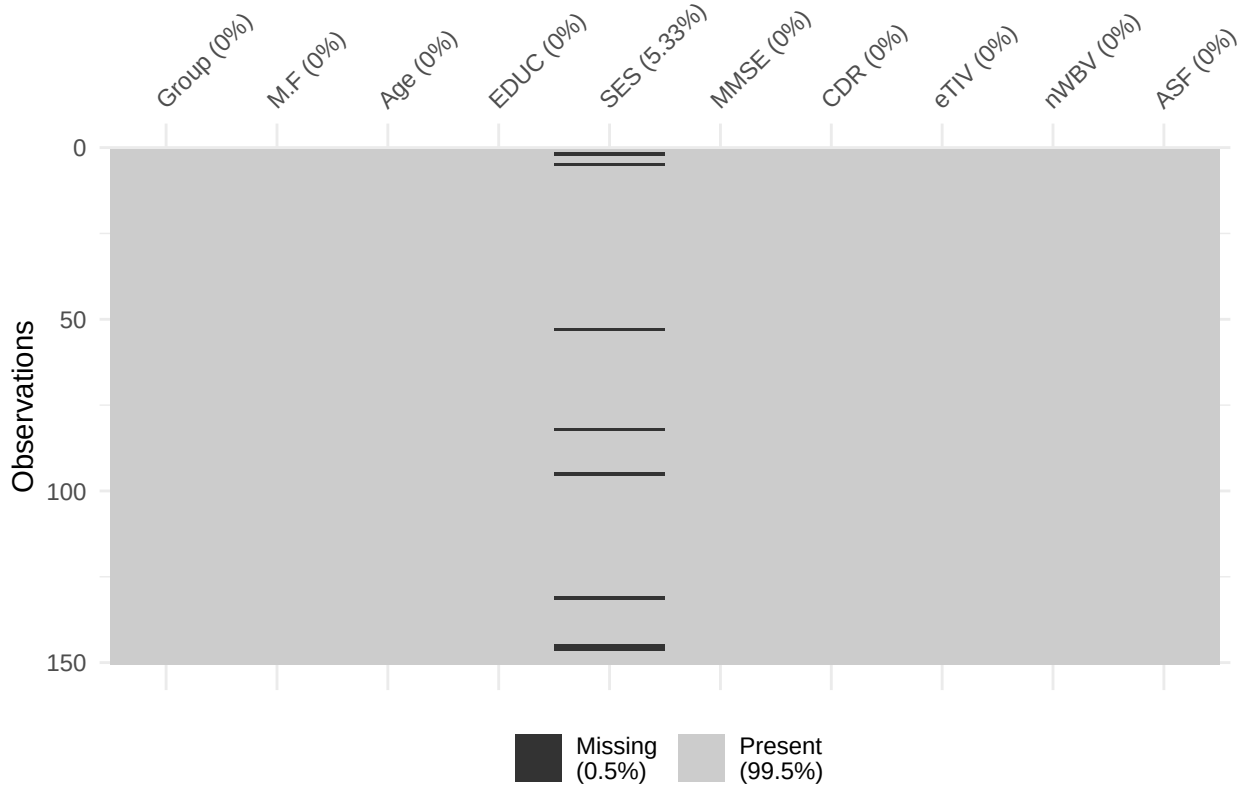
The dataset we will be working with is from the Open Access Series of Imaging Studies (OASIS), which is a compilation of open-source distributing MRI datasets. OASIS is made available by the Washington University Alzheimer's Disease Research Center. This dataset is licensed under the CCO 1.0 Universal (CCO 1.0) Public Domain Dedication and is open source where no research ethics approval is required. For the purposes of this project, we are assuming that the measurements all the measurements were taken at the end of the study.

The dataset can be obtained in a csv format that was read into R and analyzed using built-in packages in the R library. Columns for 'Hand','MRI.ID', 'Visit','Subject_ID' and, 'MR.Delay' were excluded from our study. All the patients were right hand and we are only considering the first visit for all the patients so 'Visit' and 'MR.Delay' are not needed. 'MRI.ID' and 'Subject_ID' are not informative so we excluded that column from our analysis.

Using the *vis miss* function from the visdat package in R, we created a heatmap to determine the percentage of missing data in each column. There are 8 missing values in the SES variable. Since SES cannot be determined on other values, we will remove the 8 individuals leaving us with 142 left for analysis.

There are 14 patients in our dataset who under the *Converted* group however we want to make our group variable a binary variable. We moved 14 patients in the *Converted* group into the *Demented*. Finally we will convert the categorical columns of 'Group', 'sex' and 'SES' into factors, with 'Group' having two levels (Nondemented, Demented), 'sex' having 2 levels ('F','M').

From our exploratory analysis, we noticed that for the lowest SES level (5) only had 3 patients in total (1-Nondemented, 2-Demented). As a rule of thumb, we need at least 10 observations for each unique level to provide us with reasonable estimates and standard errors. We have chosen to collapse the lowest level of SES (5) to the second lowest (4) which adjust the SES covaraite to have 4 levels with 1 being the highest and 4 being the lowest.




```
##         Group SEX Age EDUC SES MMSE CDR eTIV  nWBV   ASF
## 1 Nondemented   M  87   14   2   27 0.0 1987 0.696 0.883
## 2 Nondemented   F  88   18   3   28 0.0 1215 0.710 1.444
## 3 Nondemented   M  80   12   4   28 0.0 1689 0.712 1.039
## 4 Nondemented   F  93   14   2   30 0.0 1272 0.698 1.380
## 5    Demented   M  68   12   2   27 0.5 1457 0.806 1.205
## 6    Demented   F  66   12   3   30 0.5 1447 0.769 1.213
```

2.5 Statistical Methods

2.5.1 Primary Outcome

We will be using logistic regression to answer our research question to determine the impact SES and education have on dementia. The reason we are using logistics regression for our analysis is due to our primary outcome, dementia, will be examined as a binary outcome of no dementia and dementia.

2.5.2 Secondary Outcome

We will examine any potential effect modifier

2.5.3 Tertiary Outcome

Ordinary logisitic regression with CDR of 0 0.5 and 1.0

2.5.4 Software for Analysis

RStudio (Version 1.4.1717) was used to perform the data cleaning and statistical analysis on the relationship between dementia and age, considering potential confounding and effect modifiers.

3 Results

3.1 Descriptive Statistics

The descriptive statistics for the potential predictors of dementia (n = 70) and nondemented (n = 72) for individuals between the ages of 60 - 96 years old can be found in the figure (#) using the *gtsummary* package in R. The p-values reported in the figure are calculated for the Person's Chi-Squared test for categorical variables with expected cell counts ≥ 5, Wilcoxon rank sum test for numerical variables and, a Fisher's exact test for categorical variables with an expected cell count < 5. We used a significance level of $alpha = 0.05$ to compare our findings.

3.1.1 Demographic variables

There were significantly more female participants in the nondementia group (69%) than the dementia group (49%, p-value = 0.011). The mean age of people who did not have dementia was 75 years old (sd = 8) which is exactly the same as those in the dementia group (75 years old, sd =7; p-value = >0.9). The mean years of education were also the same in both the nondemented (15 years, sd = 3) and demented groups (14 years, sd = 3; p-value = 0.029).

We see high a higher proportion of participants with higher SES levels in the no dementia group (SES1 = 21%, SES2 = 38%) than in the dementia group (SES1 = 26%, SES2 = 21%, p-value = 0.3). While more participants in the dementia group were found to be in the lower SES levels (SES4 = 19%) compared to those in the no dementia group (SES4 = 27%, p-value = 0.3).

3.1.2 Derived Anatomic Values

The Estimated total intracranial volume (eTIV), Atlas scaling factor (ASF) and, Normalized whole brain volume (nWBV) mean values for both the non dementia group (eTIV = 1,480, sd = 184; p-value = 0.8, nWBV = 0.75, sd = 0.04; p-value = 0.002, ASF = 1.20, sd = 0.14; p-value = 0.8) and the dementia group (eTIV = 1,471, sd = 167; p-value = 0.8, nWBV = 0.73, sd = 0.03; p-value = 0.002, ASF = 1.21, sd = 0.13; p-value = 0.8).

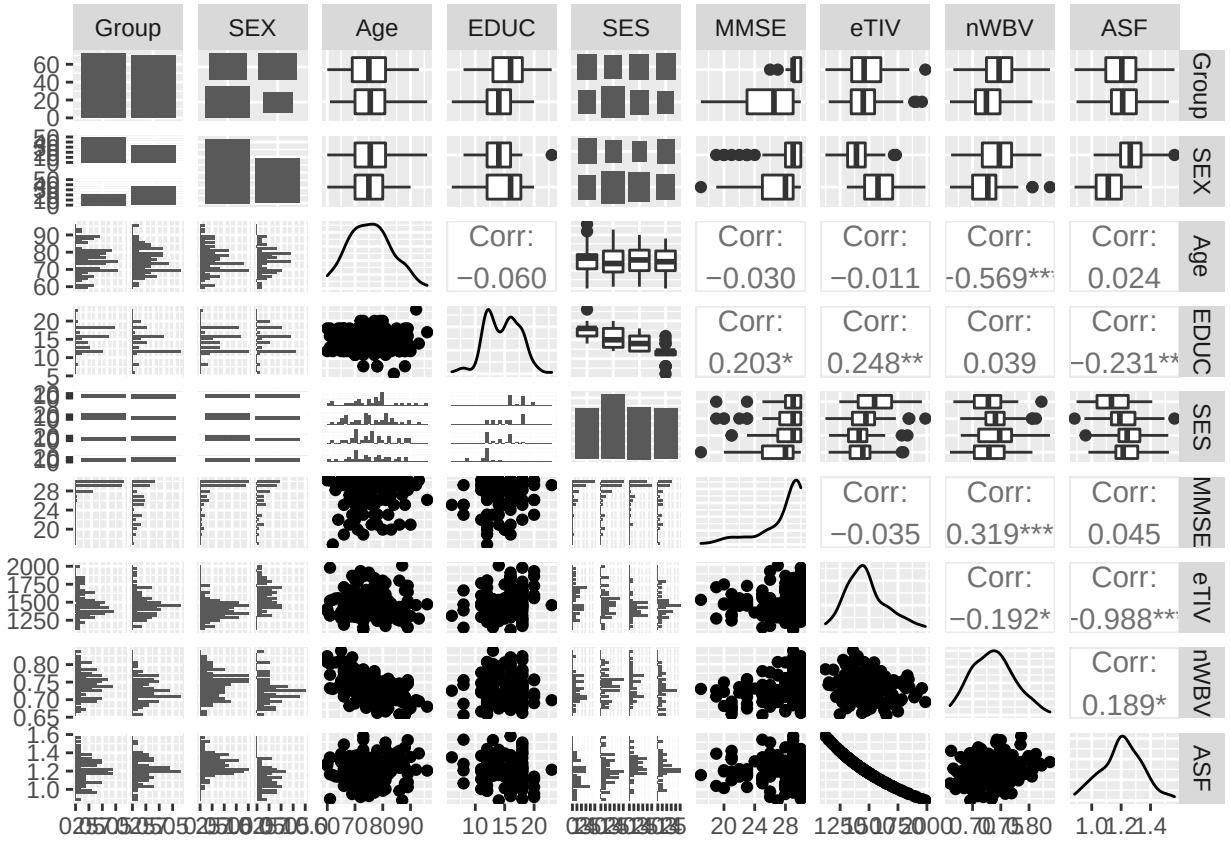
Variable	**Nondemented**, N = 72	**Demented**, N = 70	**p-value**
SEX			0.011
F	50 (69%)	34 (49%)	
M	22 (31%)	36 (51%)	

Variable	Nondemented, N = 72	Demented, N = 70	p-value
Age	75 (8)	75 (7)	>0.9
EDUC	15 (3)	14 (3)	0.029
SES			0.2
1	15 (21%)	18 (26%)	
2	27 (38%)	15 (21%)	
3	16 (22%)	18 (26%)	
4	14 (19%)	19 (27%)	
MMSE	29 (1)	26 (4)	<0.001
eTIV	1,480 (184)	1,471 (167)	0.8
nWBV	0.75 (0.04)	0.73 (0.03)	0.002
ASF	1.20 (0.14)	1.21 (0.13)	0.8

3.2 Primary Outcome Results (Logistic Regression)

3.2.1 Variable selection

We selected our variables for our final model using a pairwise plot from GGAlly, vif function from the car and the backwards stepwise variable selection from the MASS package. From the pairwise plot, there are a few variables with high correlations which could indicate multicollinearity (eTIV & ASF).



To check, we used the vif function on our full logistic model that incorporates all the predictors we

want to examine. We see that ASF and eTIV both have a VIF value that is greater than 5. As a rule of thumb, VIF values that are greater than 5 have high collinearity and need to be addressed. ASF and eTIV both have high VIF values which are above 5.

We used the Akaike information criterion (AIC) to determine which predictors will be kept in our initial model. The model that provides the lowest AIC value will be the one that provides the best fit model with the least amount of variables to avoid over and under fitting. From the backwards stepwise regression our initial AIC was 141 from the full model but after dropping ASF, we obtain a AIC of 138.27. We can check this by doing a anova test with two models, one with ASF and a reduced one without.

```
##
## Call:
## glm(formula = Group ~ SES + Age + EDUC + SEX + MMSE + eTIV +
##     nWBV + ASF, family = "binomial", data = dementia_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5376  -0.6848  -0.1148   0.6389   2.6391
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  59.141689  27.298658   2.166   0.0303 *
## SES2         -1.850064   0.698690  -2.648   0.0081 **
## SES3         -0.941306   0.787523  -1.195   0.2320
## SES4         -2.468292   0.984195  -2.508   0.0121 *
## Age          -0.096309   0.044269  -2.176   0.0296 *
## EDUC         -0.275559   0.130502  -2.112   0.0347 *
## SEXM          1.372009   0.652266   2.103   0.0354 *
## MMSE         -0.802855   0.185518  -4.328 1.51e-05 ***
## eTIV         -0.005013   0.008228  -0.609   0.5424
## nWBV        -21.504082   9.480462  -2.268   0.0233 *
## ASF          -0.940591  10.526702  -0.089   0.9288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 196.83  on 141  degrees of freedom
## Residual deviance: 118.79  on 131  degrees of freedom
## AIC: 140.79
##
## Number of Fisher Scoring iterations: 6
```

```
##           GVIF Df GVIF^(1/(2*Df))
## SES   2.808293  3        1.187792
## Age   2.175392  1        1.474921
## EDUC  2.338931  1        1.529356
```

```
## SEX   2.003545  1        1.415466
## MMSE  1.206772  1        1.098532
## eTIV 45.555634  1        6.749491
## nWBV  2.229360  1        1.493104
## ASF  44.798171  1        6.693144
```

```
##
## Call:  glm(formula = Group ~ SES + Age + EDUC + SEX + MMSE + eTIV +
##     nWBV, family = "binomial", data = dementia_sub)
##
## Coefficients:
## (Intercept)         SES2         SES3         SES4          Age         EDUC
##   57.040108    -1.857340    -0.948694    -2.477091    -0.096753    -0.277784
##        SEXM         MMSE         eTIV         nWBV
##    1.373275    -0.802380    -0.004296   -21.550511
##
## Degrees of Freedom: 141 Total (i.e. Null);  132 Residual
## Null Deviance:       196.8
## Residual Deviance: 118.8     AIC: 138.8
```

To determine if we should drop the ASF variable, we can do partial F-test to determine of the is a difference between the two models at a significance level of $\alpha = 0.05$.

Our null hypothesis is that there is a there is no difference on coefficients if we remove them from our model and the alternative is that at least one of the coefficients removed from the model is non-zero

$$H_o : \text{All coefficients removed from the model are zero}$$

$$H_a : \text{At least one of the coeffcients removed from the model is non-zero}$$

```
## Analysis of Deviance Table
##
## Model 1: Group ~ SES + Age + EDUC + SEX + MMSE + eTIV + nWBV
## Model 2: Group ~ SES + Age + EDUC + SEX + MMSE + eTIV + nWBV + ASF
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1       132     118.79
## 2       131     118.78  1 0.0080112   0.9287
```

We obtain a F test-statistic with df 1,131 and a p-value of 0.9287 which means we fail to reject the null hypothesis. This means we should drop the ASF coefficient because they do not significantly improve the fir of our model.

3.2.1.1 Confounding We expect there to be confounders in our dataset as they are linked to multiple risk factors associated with dementia. We expect age, sex, SES and, education all be confounders with dementia. From our stepwise model, the p-value for all of these co-variates are less

than our significance level of $\alpha = 0.05$, so we did not need to check for confounding by calculating the magnitude of confounding on our primary outcome.

$$\text{Magnitude of Confoudning} = \frac{|\beta_{unadjusted} - \beta adjusted}{\beta_{adjusted}}$$

Ethnicity may have a confounding effect on the incidence of AD, however, we are unable to adjust for ethnicity since it is unmeasured in our dataset.

3.2.2 Working model

Currently we have a multi-logistic model, from our stepwise selection, that is:

$$\hat{logit}(p) = \hat{\beta}_0 + \hat{\beta}_1 * \text{SES2} + \hat{\beta}_2 * \text{SES3} + \hat{\beta}_3 * \text{SES4} + \hat{\beta}_4 * \text{Age} + \hat{\beta}_5 * \text{Educ} + \hat{\beta}_6 * \text{Males} + \hat{\beta}_7 * \text{MMSE} + \hat{\beta}_8 * \text{eTIV} + \hat{\beta}_9 * \text{nWBV}$$

$$\hat{logit}(p) = 54.04 - 1.857 * \text{SES2} - 0.949 * \text{SES3} - 2.477 * \text{SES4} - 0.0968 * \text{Age} - 0.2778 * \text{Education} + 1.373 * \text{Males} - 0.802 * \text{M}$$

Where our baseline is women who are non-demented with the highest SES (SES=1).

- Outcome: Dementia Status (Binary; Nondemented/Demented)
- Main Exposure: Social economic status (Ordinal; 1,2,3,4)
- Other Variables : Age(Continuous in years; 60 - 96); Education(Continuous in years; 6 - 23); Sex(Binary; Female/Male); MMSE (Continuous; 17-30); eTIV(Continuous in mm^3; 1123 - 1987); nWBV(Continuous in mg; 0.66 - 0.8370)

```
##
## Call:
## glm(formula = Group ~ SES + Age + EDUC + SEX + MMSE + eTIV +
##     nWBV, family = "binomial", data = dementia_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5323  -0.6838  -0.1150   0.6407   2.6458
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  57.040108  13.783034   4.138 3.50e-05 ***
## SES2         -1.857340   0.694932  -2.673  0.00752 **
## SES3         -0.948694   0.783591  -1.211  0.22601
## SES4         -2.477091   0.979889  -2.528  0.01147 *
## Age          -0.096753   0.043975  -2.200  0.02779 *
## EDUC         -0.277784   0.128329  -2.165  0.03042 *
## SEXM          1.373275   0.652781   2.104  0.03540 *
```

```
## MMSE         -0.802380   0.185206  -4.332 1.48e-05 ***
## eTIV         -0.004296   0.001811  -2.372  0.01771 *
## nWBV        -21.550511   9.470697  -2.275  0.02288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 196.83  on 141  degrees of freedom
## Residual deviance: 118.79  on 132  degrees of freedom
## AIC: 138.79
##
## Number of Fisher Scoring iterations: 6
```

For this base model without effect modification we have:

- $\beta_0 = 57.040$: the log odds of dementia status of a female who with high SES when all other covariates are held at 0. However, this intercept does not make much sense as we cannot have an age or brain volumes of 0.
- β_1 = the difference in log odd of dementia status between SES of 1 and SES 2 for females.
- β_2 = the difference in log odd of dementia status between SES of 1 and SES 3 for females.
- β_3 = the difference in log odd of dementia status between SES of 1 and SES 4 for females.
- β_4 = the change in log odds of dementia status for each 1 year increase in age for females with a SES of 1.
- β_5 = the change in log odds of dementia status for each 1 year increase in education for females with a SES of 1.
- β_6 = the difference of log odds of dementia status between females and males with a SES of 1
- β_7 = the change in log odds of dementia status for ever 1 point increase in mini-mental state for females with a SES of 1.
- β_8 = the change in log odds of dementia status for ever 1 unit increase in estimated total intracranial volume for females with a SES of 1.
- β_9 = the change in log odds of dementia status for ever 1 unit increase in normalized whole brain volume for females with a SES of 1.

To convert the log odds to an odds ratio, we will need to exponentiate our beta coefficients.

- $e^{-1.856} = 0.156$ the odds ratio of SES switching from 1 to 2 is 0.156 () for
- $e^{-0.949} = 0.387$
- $e^{-2.477} = 0.0840$

- $e^{-0.0968} = 0.9078$ = For every one year increase in age, the odds of dementia is 0.9078 times larger for women in the highest SES
- $e^{-0.278} = 0.757$
- $e^{1.37} = 3.95$ = The odds of dementia for men is 3.95 times higher than for women at the highest SES level.
- $e^{-0.802} = 0.448$ = For every one point increase in MMSE, the odds of dementia is 0.488 times higher for women in the highest SES
- $e^{-0.00430} = 0.995$ = For every 1 unit increase in eTIV, the odds of dementia is 0.995 times higher for women in the highest SES.
- $e^{-21.55} = 0.00$

```
##  (Intercept)         SES2         SES3         SES4          Age         EDUC
## 5.918395e+24 1.560873e-01 3.872465e-01 8.398721e-02 9.077799e-01 7.574606e-01
##         SEXM         MMSE         eTIV         nWBV
## 3.948261e+00 4.482607e-01 9.957132e-01 4.372522e-10
```

```
##                    2.5 %       97.5 %
## (Intercept) 1.608424e+14 7.205440e+37
## SES2        3.684554e-02 5.772146e-01
## SES3        7.913428e-02 1.754295e+00
## SES4        1.111872e-02 5.395149e-01
## Age         8.287371e-01 9.861055e-01
## EDUC        5.806726e-01 9.647200e-01
## SEXM        1.139256e+00 1.511559e+01
## MMSE        2.967445e-01 6.142595e-01
## eTIV        9.919742e-01 9.991115e-01
## nWBV        1.430100e-18 2.908413e-02
```

Characteristic	OR	95% CI	p-value	GVIF	Adjusted GVIF
SES			**0.011**	2.8	1.2
1					
2	0.16	0.04, 0.58			
3	0.39	0.08, 1.75			
4	0.08	0.01, 0.54			
Age	0.91	0.83, 0.99	**0.021**	2.2	1.5
EDUC	0.76	0.58, 0.96	**0.024**	2.3	1.5
SEX			**0.030**	2.0	1.4
F					
M	3.95	1.14, 15.1			
MMSE	0.45	0.30, 0.61	**<0.001**	1.2	1.1
eTIV	1.00	0.99, 1.00	**0.013**	2.2	1.5
nWBV	0.00	0.00, 0.03	**0.019**	2.2	1.5

3.2.3 Goodness of fit

We can do a hypothesis test for the goodness of fit at the significance level of $\alpha = 0.05$, where our null hypothesis is the the model is correct and the data fits. Our alternative hypothesis is that there

is an evidence of a lack of fit.

$$H_o : \text{The data is a fits our selected model distrbution}$$

$$H_a : \text{The data is not consistent with our selected model distribution}$$

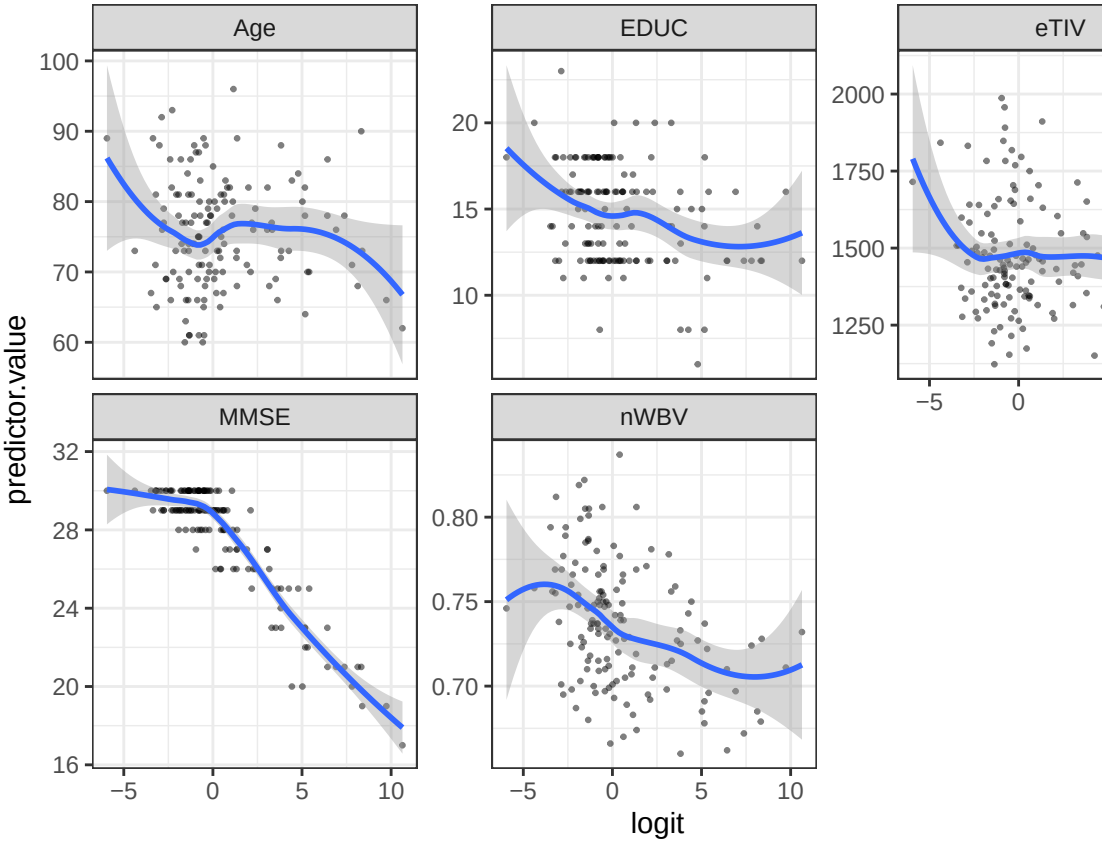
```
## [1] 0.7882735
```

With a Residual deviance of 118.79 on 132 degrees of freedom, we get a probability of 0.7882735, which is greater than our significance level of $\alpha = 0.05$. We fail to reject our null hypothesis and we can say that our data is a good fit with our selected model.

we tested our model using the following logistic regression assumptions:

3.2.4 Linearity Assumption

The first assumption for logistic regression is that the relationship between X and the log-odds is linear. We can do this visually by plotting the continuous predictors and the logit of the outcome and using the Box-Tidwell test and adding all the continuous co-varaites and their corresponding natural log. With an acceptance criteria of $\alpha = 0.05$



3.2.4.1 Scatterplots

From the graphs we see that MMSE, Education and nWBV are pretty linearly associated with the log-odds. Age does not look linear and might need to be transformed to a higher power (cubic)

3.2.4.2 Box-Tidwell The Box-Tidwell test adds interactions between the continuous variables and the corresponding natural log into the model. Our null hypothesis is that the our continuous variables are linearly related to the log odds and the alternative hypothesis is that our continuous variables are not linearly related to the log odds. We will be using a significant level of $\alpha = 0.05$.

$$H_o : \text{Continuous x variables are linearly related to the log-odds}$$

$$H_a : \text{Continuous x variables are not linearly related to the log-odds}$$

```
##                    MLE of lambda Score Statistic (z) Pr(>|z|)
## dementia_sub$Age         0.36427              0.2876   0.7736
## dementia_sub$EDUC        0.47207              0.1121   0.9108
## dementia_sub$MMSE        1.54500             -0.9509   0.3416
## dementia_sub$eTIV        1.56662             -0.1761   0.8602
## dementia_sub$nWBV       -0.30807              0.5896   0.5554
##
## iterations =  15
```

From the Box-Tidwell test, we see that the p-values for all of our continuous variables are greater than our significance level of $\alpha = 0.05$ Since the p-values are not significant for all our continuous variables, we can say that our continuous x variables are linearly related to the log-odds.

3.2.5 Multicollinearity

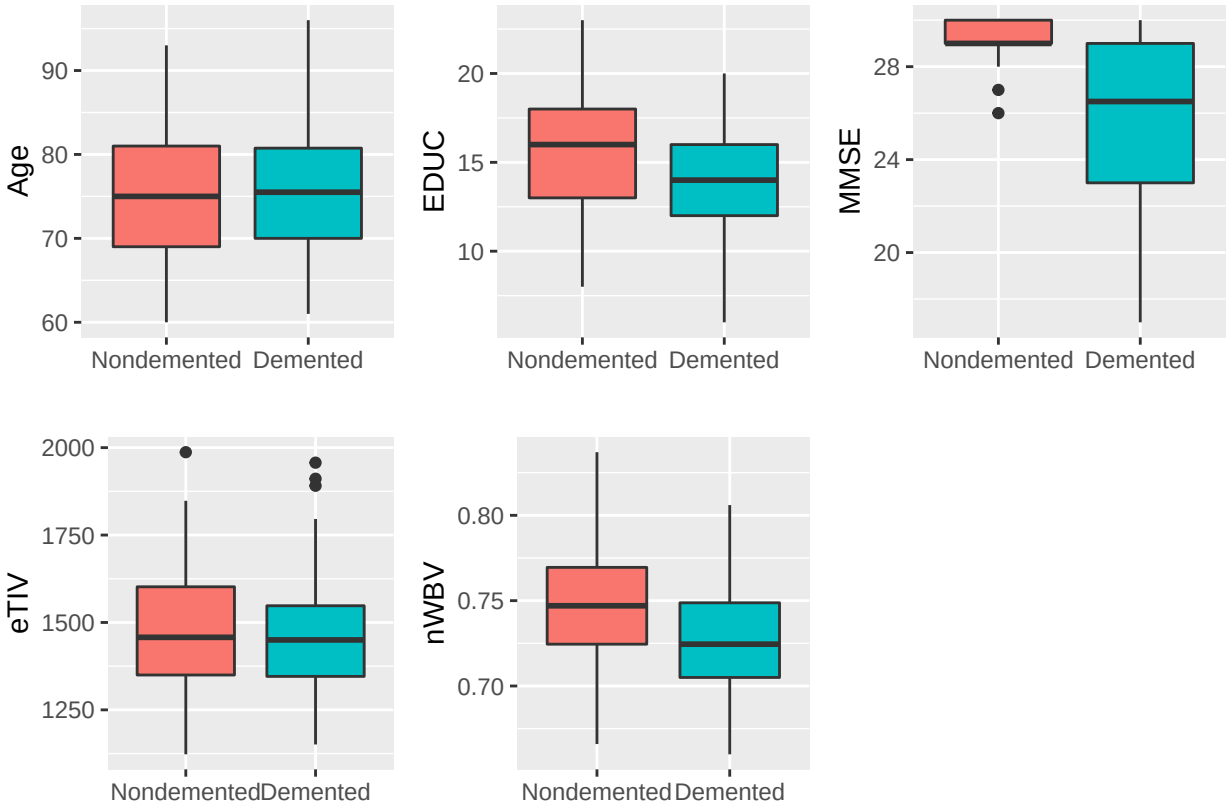
Multicollinearity was assessed earlier during during our variable selection. The presence of correlation can cause errors in our results by using redundant information. This could reduce our precision in our analysis. Running the VIF method again on our current model, we don't see any evidence of collinearity between any of our predictor variables.

```
##          GVIF Df GVIF^(1/(2*Df))
## SES  2.769844  3        1.185066
## Age  2.151668  1        1.466856
## EDUC 2.258286  1        1.502759
## SEX  2.008336  1        1.417158
## MMSE 1.206625  1        1.098465
## eTIV 2.222543  1        1.490820
## nWBV 2.226062  1        1.491999
```

3.2.6 Influential Points and Outliers

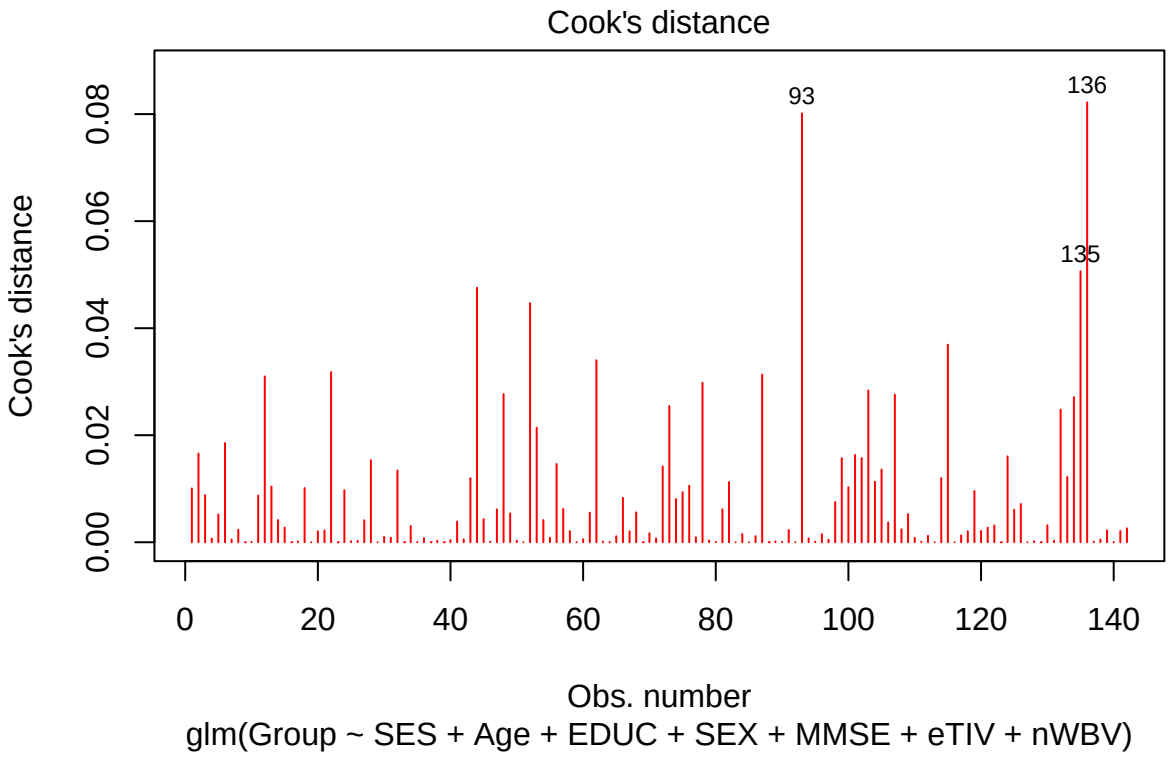
To identify any outliers and influence point in our dataset, we will use boxplots, histograms and look at Cook's distance values.

3.2.6.1 Boxplots We can visually look for outliers using boxplots on our continuous variables.



We see that MMSE and eTIV both have outliers however, not all outliers are influential observations. To check if they are influential we will look at Cook's Distance measures.

3.2.6.2 Cook's Distance As a rule of thumb, if a point has a Cook's distance that is greater than 1, then the data point is likely to be influential.



From the plot, we see that the 3 highest values are from point 93 , 135 and 136. However, all 3 points have a Cook's distance that is less than 1 so we can leave them in our model.

3.3 Secondary Outcome Results (Effect modification/Confounding)

Using all the variables, we tested for effect modification between SES and nWBV and EDUC with nWVB. To test this we will do a likelihood ratio test without and without the interaction term to determine if we need to leave them in our model. The null hypothesis is that there is no difference between the models and our alternative hypothesis is that there is a difference between the two models. We will test this at a significance level of $\alpha = 0.05$.

```
##
## Call:
## glm(formula = Group ~ (SES + Age + EDUC + SEX + MMSE + eTIV +
##     nWBV)^2, family = "binomial", data = dementia_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8977  -0.3040   0.0000   0.1862   1.9917
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.772e+02  6.394e+02   1.372   0.1701
```

```
## SES2         2.391e+02  2.286e+02   1.046   0.2957
## SES3         3.172e+00  7.117e+01   0.045   0.9644
## SES4        -1.033e+02  1.569e+02  -0.658   0.5104
## Age         -1.603e+00  2.312e+00  -0.693   0.4882
## EDUC        -2.429e+01  2.413e+01  -1.007   0.3142
## SEXM        -2.509e+01  8.311e+01  -0.302   0.7627
## MMSE        -4.953e+00  1.552e+01  -0.319   0.7497
## eTIV        -1.199e-01  2.382e-01  -0.503   0.6148
## nWBV        -1.121e+03  6.773e+02  -1.656   0.0978 .
## SES2:Age    -1.913e-01  3.889e-01  -0.492   0.6227
## SES3:Age    -1.715e-01  2.504e-01  -0.685   0.4933
## SES4:Age     1.123e-03  5.866e-01   0.002   0.9985
## SES2:EDUC   -3.058e+00  2.043e+00  -1.497   0.1344
## SES3:EDUC   -3.409e-01  7.593e-01  -0.449   0.6534
## SES4:EDUC   -2.084e-01  1.033e+00  -0.202   0.8402
## SES2:SEXM    2.100e+01  1.316e+01   1.595   0.1106
## SES3:SEXM    4.687e+00  4.853e+00   0.966   0.3342
## SES4:SEXM    9.792e+00  8.092e+00   1.210   0.2262
## SES2:MMSE   -4.997e+00  4.063e+00  -1.230   0.2188
## SES3:MMSE   -4.362e-01  1.141e+00  -0.382   0.7022
## SES4:MMSE   -3.319e+00  2.715e+00  -1.222   0.2216
## SES2:eTIV   -3.254e-02  2.454e-02  -1.326   0.1849
## SES3:eTIV   -4.360e-03  1.354e-02  -0.322   0.7474
## SES4:eTIV   -3.637e-03  2.074e-02  -0.175   0.8608
## SES2:nWBV    1.483e+00  1.345e+02   0.011   0.9912
## SES3:nWBV    4.304e+01  7.186e+01   0.599   0.5492
## SES4:nWBV    2.710e+02  1.460e+02   1.856   0.0634 .
## Age:EDUC     5.535e-02  8.099e-02   0.683   0.4943
## Age:SEXM     1.294e-01  3.215e-01   0.403   0.6873
## Age:MMSE    -8.785e-03  5.306e-02  -0.166   0.8685
## Age:eTIV    -3.969e-04  7.099e-04  -0.559   0.5761
## Age:nWBV     1.913e+00  2.143e+00   0.893   0.3721
## EDUC:SEXM    1.572e+00  9.718e-01   1.618   0.1057
## EDUC:MMSE   -2.129e-01  2.919e-01  -0.729   0.4659
## EDUC:eTIV    2.057e-04  2.738e-03   0.075   0.9401
## EDUC:nWBV    3.417e+01  1.909e+01   1.790   0.0734 .
## SEXM:MMSE    4.758e-01  1.060e+00   0.449   0.6534
## SEXM:eTIV    7.766e-03  1.028e-02   0.756   0.4499
## SEXM:nWBV   -4.850e+01  8.116e+01  -0.598   0.5501
## MMSE:eTIV    1.779e-03  3.556e-03   0.500   0.6169
## MMSE:nWBV    7.384e+00  1.423e+01   0.519   0.6037
## eTIV:nWBV    1.163e-01  2.307e-01   0.504   0.6143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 196.826  on 141  degrees of freedom
```

```
## Residual deviance:  69.942  on  99  degrees of freedom
## AIC: 155.94
##
## Number of Fisher Scoring iterations: 11
```

```
##
## Call:
## glm(formula = Group ~ SES + Age + EDUC + SEX + MMSE + eTIV +
##     nWBV + nWBV * SES, family = "binomial", data = dementia_sub)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.44938  -0.61232  -0.09122   0.52467   2.47430
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  65.838830  18.368476   3.584 0.000338 ***
## SES2          1.259946  14.302324   0.088 0.929802
## SES3         12.908751  14.982198   0.862 0.388904
## SES4        -31.275069  15.347700  -2.038 0.041573 *
## Age          -0.113318   0.048206  -2.351 0.018737 *
## EDUC         -0.310590   0.136384  -2.277 0.022768 *
## SEXM          1.332826   0.693998   1.921 0.054794 .
## MMSE         -0.919432   0.212470  -4.327 1.51e-05 ***
## eTIV         -0.004224   0.001924  -2.195 0.028143 *
## nWBV        -26.606769  15.946647  -1.668 0.095219 .
## SES2:nWBV    -4.282035  19.499143  -0.220 0.826182
## SES3:nWBV   -18.474619  20.221142  -0.914 0.360912
## SES4:nWBV    38.667193  20.604286   1.877 0.060565 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 196.83  on 141  degrees of freedom
## Residual deviance: 110.14  on 129  degrees of freedom
## AIC: 136.14
##
## Number of Fisher Scoring iterations: 6
```

```
## Analysis of Deviance Table
##
## Model 1: Group ~ SES + Age + EDUC + SEX + eTIV + nWBV
## Model 2: Group ~ SES + Age + EDUC + SEX + MMSE + eTIV + nWBV + nWBV *
##     SES
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       133     163.03
```

```
## 2       129     110.14  4   52.898 8.952e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4 Tertiary Outcome Results (Subgroup Analysis)

Demographic stochasticity refers to fluctuations in population sizes or densities that arise from the fundamentally discrete nature of individual birth and death events. Demographic stochasticity is a particularly instructive case for illustrating a mechanism for how noise arises as an aggregate description from a lower-level mechanistic process. We summarize the myriad lower-level processes that mechanistically lead to the event of a 'birth' in the population as a probability: In a population of N identical individuals at time t, a birth occurs with probability b_t(N_t) (*i.e.* a rate that can depend on the population size, N), which increases the population size to N+1. Similarly, death events occur with probability d_t(N_t), decreasing the population size by one individual, to N-1. Assuming each of these events are independent, this is a state-dependent Poisson process. The change in the probability of being in state N is given by the sum over the ways to enter the state, minus the ways to leave the state: a simple expression of probability balance known as the master equation (Kampen 2007). Note that in general this approach is equally applicable to stochastic transitions of any sort, not just step sizes of +/- 1 and not just birth and death events, but can include transitions between stage classes or trait values, including mutations to continuously-valued traits in evolutionary dynamics (e.g. Boettiger et al. 2010).

4 Conclusion or Discussion

The Gillespie (1977) provides an exact algorithmcfor simulating demographic stochasticity at an individual level.

5 Contribution statement

The algorithm is a simple and direct implementation of the master equation, progressing in random step sizes determined by the waiting time until the next event. Free from both the approximations and mathematical complexity, the Gillespie algorithm is an interesting example of where we rely on a numerical implementation to check the accuracy of an analytic approximation, even in the case of simple models such as we will discuss. Though the algorithm is often maligned as numerically demanding, it can be run much more effectively even on large models on today's computers than when it was first developed in the 70s, and remains an underutilized approach for writing simple and approximation-free[1] stochastic ecological models.

6 Table and Figures

As our objective is to tie the origins of noise more closely to biological processes, it will be helpful to make the notion of a master equation concrete with a specific example. We will focus on the

[1] that is, free from the approximation made by SDE models as we see in the van Kampen example. All models are, of course, only approximations.

classic case of Levins (1969) patch model, to illustrate the Gillespie algorithm and the van Kampen system size expansion

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \underbrace{cn\left(1 - \frac{n}{N}\right)}_{\text{birth}} - \underbrace{en}_{\text{death}}, \tag{1}$$

7 Appendix

where n individuals compete for a finite number of suitable habitats N. Individuals die a constant rate e, and produce offspring at a constant rate c who then have a probability of colonizing an open patch that is simply proportional to the fraction of available patches, 1 - n/N.

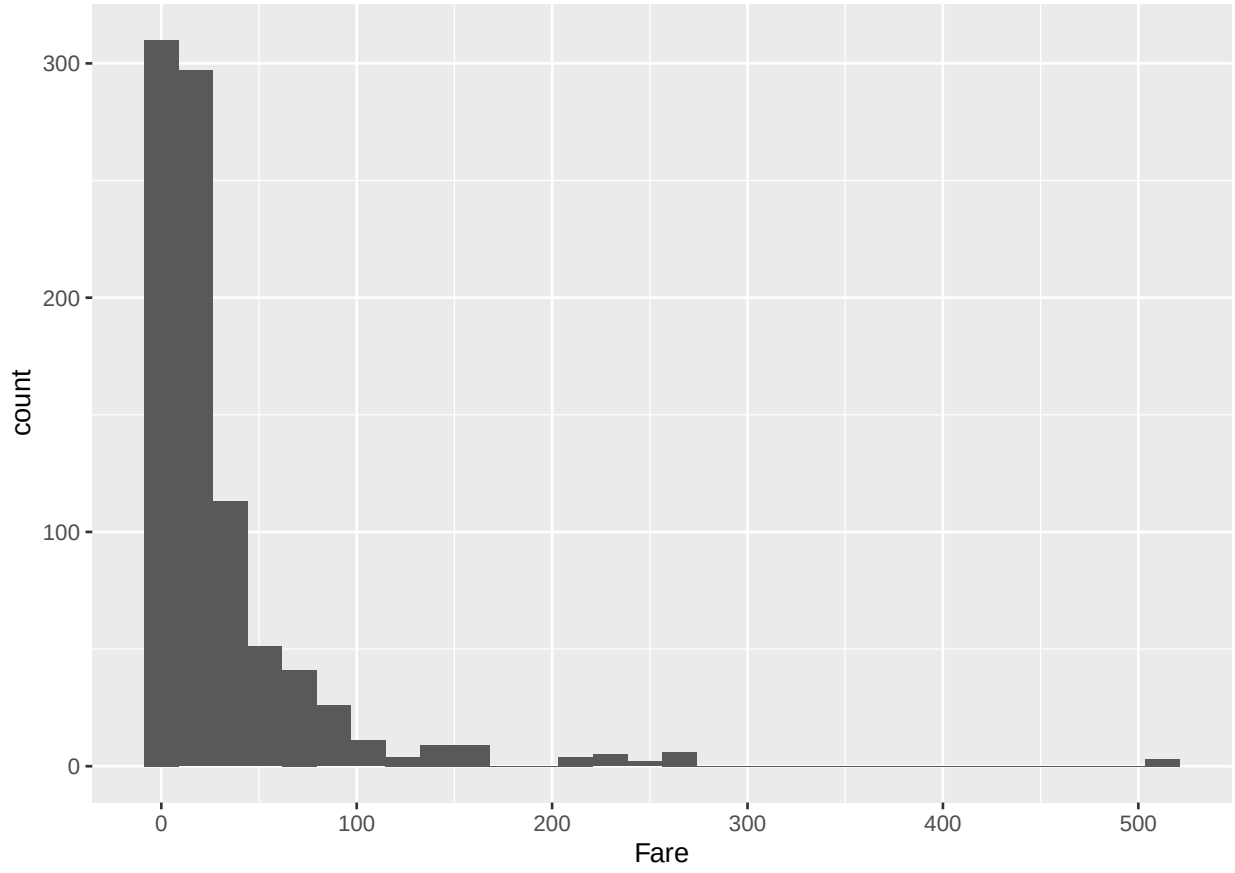


Figure 1 shows the results of two exact SSA simulations of the classic patch model of Levins (1969).

8 Conclusions

This review has explored three paradigms in how noise is viewed throughout the ecological literature, which I have dubbed respectively: noise the nuisance, noise the creator, and noise the informer. Noise can be seen as a nuisance almost by definition: in examining the origins of noise, we have seen how

stochasticity is introduced not because ecological processes are random in some fundamental sense, but rather, because those processes are influenced by a complex combination of forces we do not model explicitly. In this view, noise captures all that additional variation that is separate from the process of interest, and a rich array of statistical methods allow us to separate the one from the other in observations and experiments. By examining the origins of noise, we have seen that despite the complex ways in this noise can enter a model, that a Gaussian white-noise approximation (Kampen 2007; Black & McKane 2012) is often appropriate given a limit of a large system size – a fact often invokedn implicitly but rarely derived explicitly from the theorems of Kurtz (1978) and others.

In this context, noise does not act to create phenomena of interest directly. The sudden transitions we seek to anticipate are still explained by the deterministic part of the model – bifurcations. But nor is noise a nuisance that merely cloaks this deterministic skeleton from plain view: rather, it becomes a novel source of information that would be inaccessible from a purely deterministic approach. I believe more examples of how noise can inform on underlying processes is possible, but will require greater dialog between these world views.

9 Acknowledgements

The author acknowledges feedback and advice from the editor, Tim Coulson and two anonymous reviewers. This work was supported in part by USDA National Institute of Food and Agriculture, Hatch project CA-B-INS-0162-H.

10 References

11 Figure 1: Population dynamics from a Gillespie simulation of the Levins model with large (N=1000, panel A) and small (N=100, panel B) number of sites (blue) show relatively weaker effects of demographic noise in the bigger system. Models are otherwise identical, with e = 0.2 and c = 1 (code in appendix A). Theoretical predictions for mean and plus/minus one standard deviation shown in horizontal re dashed lines.