Data Analysis Report - Trends in the Number of PhDs Awarded From 2008 to 2017 by Major Field

Ava Gordon,

Abstract

Our project – Trends in the Number of PhDs awarded from 2008 to 2017 by Major Field" aims to answer the question "Has the number of PhD's awarded changed over time with respect to the year (2008 to 2017) and/or specific major field?". It was first necessary to decide if we should look at the count of PhD's at a broad field focus, major field focus, or an even more granular field focus. Through AIC, exploratory data analysis, and the context of our data, we found that looking at the counts of PhD's at a major field level made the most sense. Next, we constructed a model based on the Poisson GLM, but because we found evidence of over-dispersion in our model, we moved to the Quasi-Poisson to help mitigate this over-dispersion. Given that an interaction effect did not help our model, we resorted to our final model using 'year' and 'major_field' as explanatory variables (where 'major_field' is a factor variable) and response variable 'n_phds'. Almost all model terms were significant in our model summary, and we found that certain major fields showed larger increases in count than others. Overall, we discovered that while year does not play a huge role in the number of PhD's awarded, there are trends in certain major fields in regard to the count of PhD's awarded from 2008 to 2017. We invite you to take a closer look at our investigation of our research question.

Introduction

Our project goal is to discover if there are relationships in our data set that correspond to the number of PhD's (n_phds) awarded. More specifically, our question of interest is "Has the number of PhD's awarded changed over time with respect to the year (2008 to 2017) and/or specific field?". Our data set contains three field variables – broad_field (the parent field), major_field (a sub-field below the broad field), and field (the most detailed field) – so, we will investigate which of those three variables is most influential in determining the number of PhD's awarded, with respect to the context of our data and question of interest to determine what model we will go with. Overall, we wish to discover what factors are influential in the count of PhD's awarded. Through doing so, we may learn insights that

could be useful for individuals who wish to pursue a PhD, or are just curious about these results in general.

Data and Methods

The data we used in our analysis was pulled from the National Science Foundation (NSF), an independent agency apart of the U.S federal government that supports national sciences and engineering. They produced a report in 2017 where they surveyed individuals who received doctorate degrees from accredited U.S academic institutions. This report contains 72 different assembled data sets including the one we are analyzing. See the full report Foundation (2017) for further details.

The data set that we chose for this report was cleaned by Tom Mock where he removed duplicates and created separations between the following fields: broad_field, major_field, and field (see Mock and Foundation (2019)). There are five total variables in this data-frame including the three aforementioned fields as well as 'year' and 'n_phds'. Additionally, there are 3,370 rows of data.

broad_field (character): this is the parent field (highest delineator) and includes seven categories. These categories include "Life sciences", "Mathematics and computer sciences", "Psychology and social sciences", "Engineering", "Education", "Humanities and arts", and "Other".

major_field (character): this is the sub-field below broad_field and includes 25 categories.

field (character): this represents the finest field and the most detailed. This includes 336 categories.

year (integer): this represents the year the PhD was awarded. The years included are 2008 to 2017.

n_phds (double): this represents the total number of PhDs awarded. There are 558 different total number of PhDs awarded.

Our method for observing the impacts of year and different fields on the number of PhD's awarded will be to conduct a Poisson regression. Poisson regressions can be used to model the mean counts of data for different categories. In this case, we are observing the counts of PhDs awarded yearly to certain fields.

In determining which variables to include, model comparison was conducted. We started by observing three Poisson regressions for 'broad_field', 'major_field', and 'field' with 'year'. We understand that none of these three factors can be included in the same model as they are all correlated and associated with each other as they simply just get more specific. The AIC was noticed to decrease as the fields got more specific, however, it was decided that using 'field' as a co-variate created a model that was too granular and would result in our model having over 336 coefficients in our model, despite having the lowest AIC. We were now concerned with models including 'broad_field' and 'year', along with 'major_field' and

'year' where 'broad_field' includes 7 fields and 'major_field' includes 25 fields, which is much more reasonable than 336 fields. 'Major_field' and 'year' had the smaller AIC.

We then observed a Quasi-Poisson model for these two models, and the dispersion parameter was much greater than 1 for both models, suggesting that the variance in the counts is much greater than the mean in the counts (variance increases faster than the mean -> over-dispersion), so a Quasi-Poisson model must be used. AIC cannot be determined for Quasi-Poisson models, but we decided to use the Quasi-Poisson model for 'major_field' and 'year' as it offers slightly more information than 'broad_field' (25 fields instead of 7) and had a slightly less dispersion parameter, which could help with our results. We also decided that we will still be able to observe trends in the number of PhDs awarded for broad fields through our major fields model.

Once deciding on a Quasi-Poisson model including 'year' and 'major_field', we looked into including the interaction term year*major_field. The introduction of this interaction term added several parameters and reduced the number of significant terms, so we determined that a Quasi-Poisson model including 'major_field' and 'year' is the model that would provide the most insight for our data.

This model will allow us to observe the impact of year on the number of PhD's awarded as well the difference in number of PhD's awarded among different major fields. We will observe the model intercept, coefficients, and their significance to determine the impact of these factors on the number of PhD's awarded. Since we are using a Quasi-Poisson, t-tests will be used to determine the significance of the parameters.

Our statistical model is as follows:

Let Y_i be the number of PHD's awarded to major field i, where $i = 1, 2, ..., 25$:

```
# A tibble: 25 x 3
     Y_i major_field                                n_phds
   <int> <chr>                                     <dbl>
 1     1 Agricultural sciences and natural resources 13125
 2     2 Anthropology                                 5074
 3     3 Biological and biomedical sciences          85637
 4     4 Business management and administration      14672
 5     5 Chemistry                                   25015
 6     6 Communication                                6340
 7     7 Computer and information sciences           18395
 8     8 Economics                                   11752
 9     9 Education administration                    12428
10    10 Education research                          25577
# i 15 more rows
```

Y_i follows a Quasi-Poisson distribution with $E[Y_i] = \lambda_i$, and $Var(Y_i) = \phi\lambda_i$.
$log(\lambda_i) = \beta_0 + (\beta_i)^{major} + \beta_{25} * year$, where $(\beta_1)^{major} = 0$.
This means that the reference level for this model is the major 'Agricultural Sciences and Natural Resources'.

Results:

Model summary:

```
MODEL INFO:
Observations: 3092
Dependent Variable: n_phds
Type: Generalized linear model
  Family: quasipoisson
  Link function: log

MODEL FIT:
²(25) = 200609.44, p = 0.00
Pseudo-R² (Cragg-Uhler) = 1.00
Pseudo-R² (McFadden) = 0.26
AIC =  NA, BIC =  NA

Standard errors: MLE
```

	Est.	S.E.	t val.	p
(Intercept)	12.94	17.11	0.76	0.45
major_fieldAnthropology	1.67	0.29	5.80	0.00
major_fieldBiological and biomedical sciences	1.45	0.16	8.86	0.00
major_fieldBusiness management and administration	0.73	0.21	3.51	0.00
major_fieldChemistry	1.70	0.19	9.04	0.00
major_fieldCommunication	0.68	0.27	2.56	0.01
major_fieldComputer and information sciences	2.32	0.20	11.65	0.00
major_fieldEconomics	2.14	0.22	9.67	0.00
major_fieldEducation administration	1.93	0.22	8.84	0.00
major_fieldEducation research	1.37	0.19	7.32	0.00
major_fieldForeign languages and literature	0.20	0.27	0.77	0.44
major_fieldGeosciences, atmospheric sciences, and ocean sciences	0.54	0.20	2.78	0.01
major_fieldHealth sciences	1.01	0.19	5.23	0.00
major_fieldHistory	0.82	0.23	3.61	0.00
major_fieldLetters	1.01	0.21	4.87	0.00
major_fieldMathematics and	0.98	0.20	4.85	0.00

```
statistics
major_fieldNon-S&E fields          0.61    0.25     2.45   0.01
nec
major_fieldOther education         1.06    0.32     3.28   0.00
major_fieldOther                   0.48    0.20     2.41   0.02
engineering
major_fieldOther humanities        0.88    0.19     4.51   0.00
and arts
major_fieldOther social            0.72    0.20     3.59   0.00
sciences
major_fieldPhysics and             1.52    0.17     8.73   0.00
astronomy
major_fieldPsychology              2.12    0.16    13.00   0.00
major_fieldTeacher                -0.11    0.43    -0.25   0.80
education
major_fieldTeaching fields         0.09    0.24     0.39   0.70
year                              -0.00    0.01    -0.52   0.60
-------------------------------------------------------------

Estimated dispersion parameter = 303.35
```

This model gives coefficients for each major_field that describe the change in the log of the number of PHD's awarded in reference to the major Agricultural Sciences and Natural Resources, assuming year is kept constant. These coefficients allow us to determine the differences in the number of PHD's awarded bases on the major fields. There is also a coefficient for year that describes the change in the number of PHDs as the year increases from 2008 to 2017. The dispersion parameter is 303.35, meaning that $Var(\lambda_i) = 303.35 * \lambda_i$. This means that the number of PHDs awarded does not closely follow a Poisson as the variance is much larger than the mean, supporting our use of a Quasi-Poisson model.

It took 6 Fisher Scoring iterations to estimate the maximum likelihood estimators $\hat{\beta}_0, ...\hat{\beta}_{25}$.

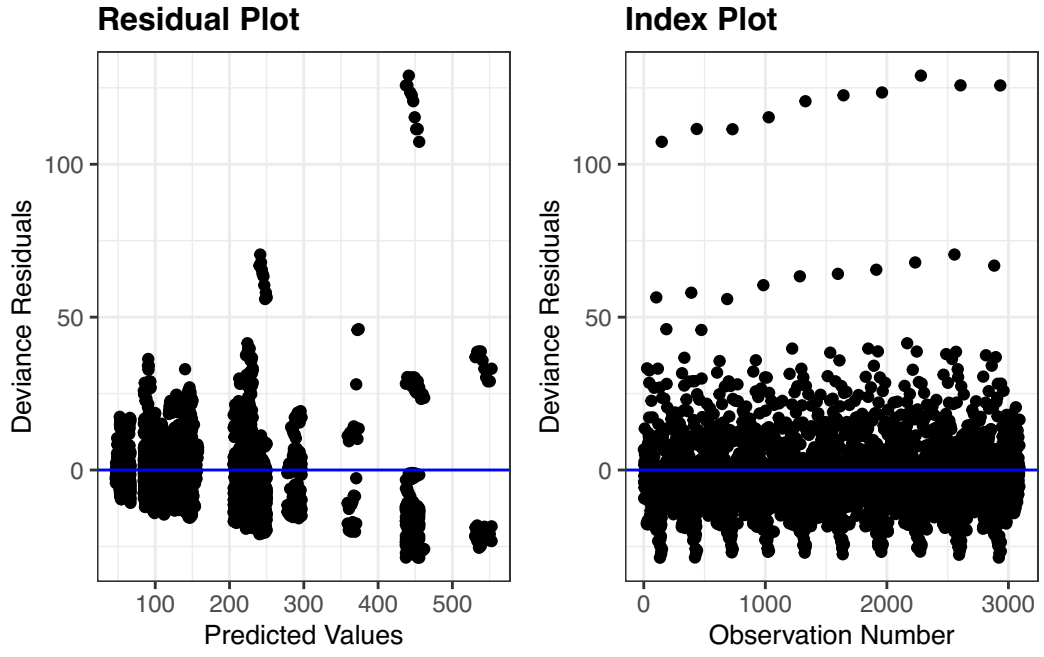
At a significance level of .05, almost all of the model terms from the coefficient summary are significant, excluding $\hat{\beta}_{10}$ ~ Foreign Languages and Literature, $\hat{\beta}_{23}$ ~ Teacher Education, $\hat{\beta}_{24}$ ~ Teaching Fields, and $\hat{\beta}_{25}$.

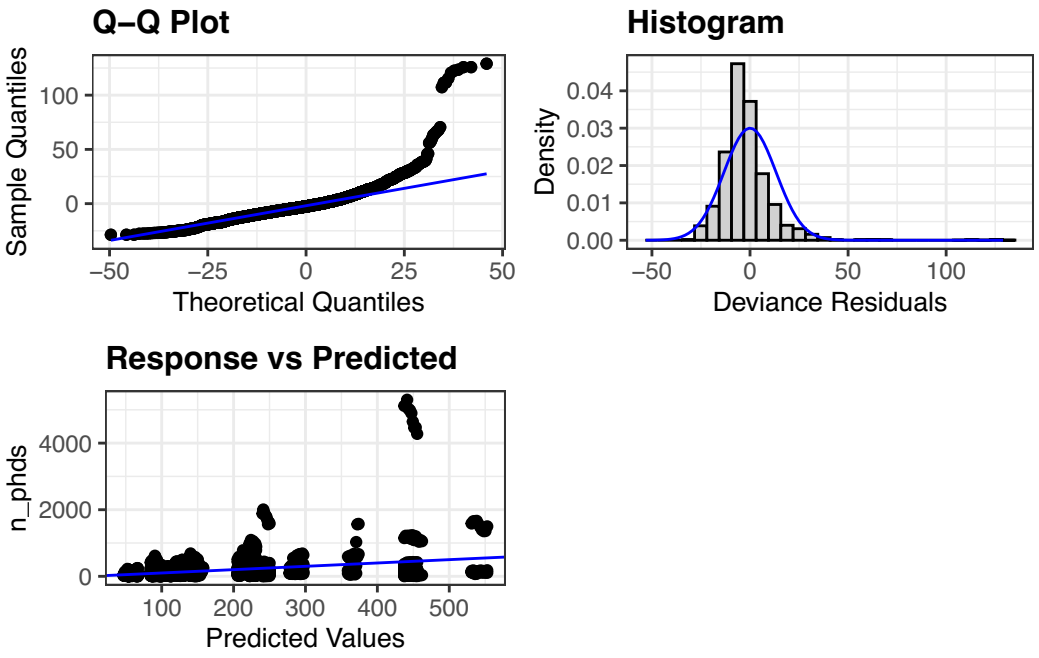
Keeping 'year' constant, the incidence of almost all major fields increases, with respect to 'Agricultural Sciences and Natural Resources'. For example, $e^{2.12} = 8.33113748$ is the estimated multiplicative change in the count of 'Psychology' PhD's for a one year increase with reference to the reference level, 'Agricultural Sciences and Natural Resources'. As another example, $e^2.32 = 10.17567430$ is the estimated multiplicative change in the count of 'Computer and information sciences' PhD's for a one year increase with reference to the reference level, 'Agricultural Sciences and Natural Resources'.'Psychology', 'Economics' and 'Computer and information sciences' are the top three major fields in the model summary that we see the largest increases in incidences. We should further investigate why this is in the context of our research question.

On the other hand, also keeping 'year' constant, the incidence of the major field 'Teacher education' decreases with respect to 'Agricultural Sciences and Natural Resources'. $e^{-0.11} = 0.89583413$ is the estimated multiplicative change in the count of 'Teacher education' PhD's for a one year increase with reference to the reference level, 'Agricultural Sciences and Natural Resources'. It makes some intuitive sense why we don't see large numbers of PhD's in 'Teacher education' as to be a teacher at a K-12 level, a PhD is not required. There are likely few people who go on to teach teaching at the university level and need a PhD in order to do so.

A smaller deviance is evidence of a better fitting model, so because the residual deviance 556945 is less than the null deviance 757555, we have reason to believe that the full model is a better fit than the reduced/NULL model with just β_0: the intercept-only model.

Residual plot and QQplot:





If the model fits well, the deviance residuals should be approximately normal with mean zero and unit variance. These residual plots suggest that our model is highly variable. The residuals do not appear to be centered around zero due to outliers, therefore, our results should be closely examined. We can see that our model fails to account for some significant variability in the Quasi-Poisson counts. The normal QQ-plot shows that a normal assumption is somewhat reasonable, but is an imperfect fit for the residuals. Therefore, our results should be closely examined and interpreted within the context of our data and analysis of our findings.

Conclusion

Using our models, we were successfully able to find relationships in our data with the number of PhD's awarded. Specifically, the major field of 'Psychology' had the largest increase in incidence. We can infer that Psychology is popular in some regard because pursuing a PhD in Psychology is recommended or sometimes necessary for most psychology related practices. Craighead and Craighead (2006) goes into more detail on the structure of Psychology PhD training.

Another conclusion we came to was the insignificance of year in relation to number of PhD's awarded. This is most likely due to the closeness in years where trends would not be obvious when it comes to PhD's awarded. However, it might be beneficial for future research to gather data from the past 6 years, especially after the COVID-19 pandemic where universities and students were heavily affected. Other trends and significance in the year awarded might be present in more recent data.

Figures and Tables

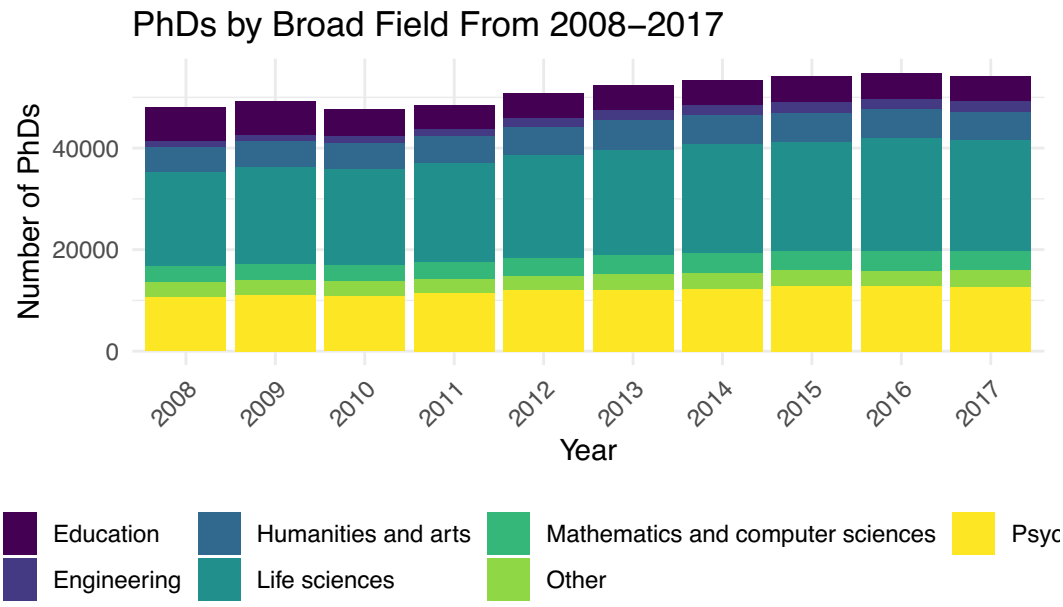


Fig 1: The Broad Field distribution remains relatively constant each year

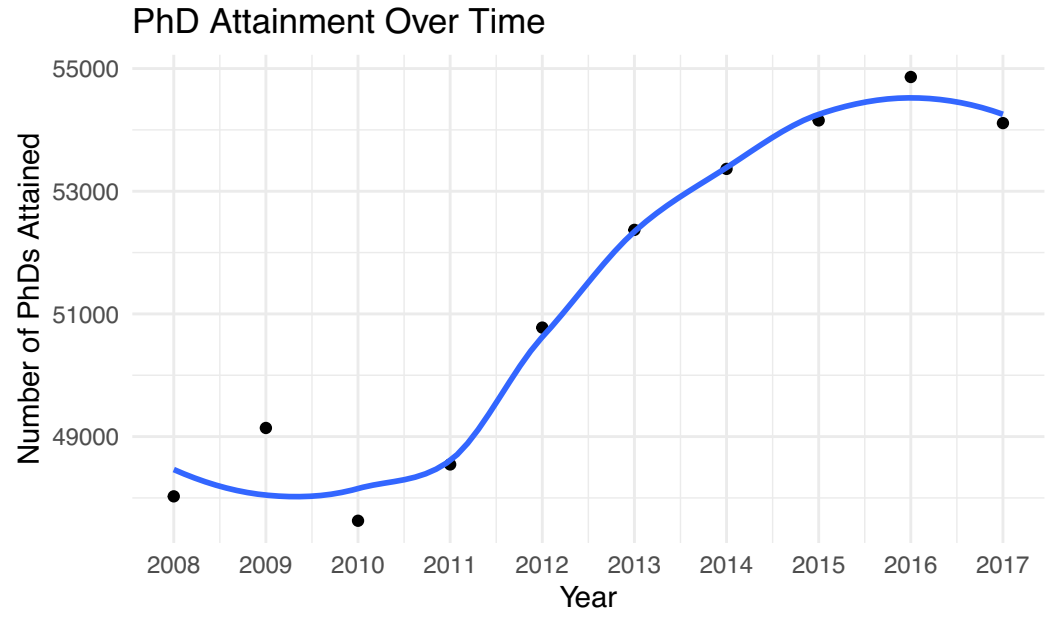


Fig 2: The number of PhD's attained per year generally increases over time

Number of PhDs in Each Major Field Grouped By Broad Field

Major Field	Broad Field PhDs	Major Field PhDs
Life sciences		
Biological and biomedical sciences	1236	378
Agricultural sciences and natural resources	1236	246
Geosciences, atmospheric sciences, and ocean sciences	1236	216
Physics and astronomy	1236	168
Health sciences	1236	142
Chemistry	1236	86
Humanities and arts		
Other humanities and arts	452	160
Letters	452	104
Foreign languages and literature	452	98
History	452	90
Psychology and social sciences		
Psychology	394	190
Other social sciences	394	160
Economics	394	26
Anthropology	394	18
Education		
Teaching fields	374	154
Education research	374	122
Teacher education	374	40
Education administration	374	34
Other education	374	24
Other		
Business management and administration	272	132
Non-S&E fields nec	272	80
Communication	272	60
Engineering		
Other engineering	210	210
Mathematics and computer sciences		
Mathematics and statistics	154	120
Computer and information sciences	154	34

Fig 3: Life Sciences are by far the most popular broad field for PhDs. Note: Large numbers are highlighted in green while smaller numbers are highlighted in purple

References

Craighead, Linda W, and W Edward Craighead. 2006. "PhD Training in Clinical Psychology: Fix It Before It Breaks." *Clinical Psychology: Science and Practice* 13 (3): 235–41.

Foundation, National Science. 2017. "Doctorate Recipients from u.s Universities." https://ncses.nsf.gov/pubs/nsf19301/.

Mock, Tom, and National Science Foundation. 2019. "PhDs Awarded by Field." https://github.com/rfordatascience/tidytuesday/tree/ master/data/2019/2019-02-1.