

Final Report: Longitudinal Analysis on the Proportion of Gender in Technology Workforce Data from Annual Stack Overflow Surveys 2014-2018

Submitted to:

Prof. Tsung-Hua Lin
Department of Statistics
University of California, Irvine

Report Prepared by:

Yuxin Fang
Project Analyst, Dept. of Statistics
University of California, Irvine

Sacha R. Uritis
Project Leader, Dept. of Statistics
University of California, Irvine

Sebastian C. Waz
Project Statistician, Dept. of Statistics
University of California, Irvine

December 13, 2018

Table of Content

Abstract	2
1 Introduction	2
2 Methodology	3
3 Results	10
4 Discussion	13
5 Conclusion	15
6 Acknowledgements	15
7 References	16
Appendix	17

Abstract

Gender bias in the technology field is irrefutable. Despite many attempts to improve diversity in this field by many organizations, research on estimating gender bias magnitude or trend over time is seriously lacking. Today, there is more recognition of various gender identities that scale beyond the traditional binary classification of male and female. In our study, we attempt to understand the gender presence in tech over the last 5 years using data from the most popular worldwide software developer survey in the modern tech community. For our longitudinal analysis, we fit a general linear mixed models with logit link with assumed banded autoregressive correlation structure using bayesian techniques. We found a significant marginal decrease in proportion of males over time {95% credible interval: (-0.190, -0.044) }. However, when we looked into the individual trajectories of our sample of countries, most countries had inconclusive results for their slopes. There were only 8 countries that showed significant decreases in proportion of males over time. We emphasize that our conclusions should be taken with speculation due to the limitations commonly found in voluntary surveys. However, we hope our study will fuel more interest and inspire research on this topic.

1 Introduction

The U.S. Department of Labor (DOL) reports that in 1950 women constituted 29.6% of the civilian workforce in the United States.¹ This percentage grew consistently over the following four decades. By 2016, the women's share of the workforce had risen to 46.8% with some of the upward momentum tapering off in the two preceding decades. Of course, the marginal share of the workforce comprised of women does not provide insight into the representation of women in specific occupations. Additionally, the trajectory of women's representation in newly emerging fields may not be adequately characterized in the workforce generally over the past several decades. A more detailed investigation is necessary.

In 2017, the U.S. Bureau of Labor Statistics (BLS) reported that 25.5% of all employees in computer and mathematical occupations in the United States were women.² This was up from 24.7% in 2015 but on par with the results in 2016 (also 25.5%).³ In recent years, the tech sector in the U.S. has developed rapidly, and it is difficult to assess whether such annual changes in women's representation may be attributed to a coherent trend in employment behavior or noise (e.g., due to volatility in the industry). The precision of an estimate for such a trend, if it exists, may be maximized by modeling correlated data on workforce representation across countries globally.

It should also be noted that the representation of gender non-conforming individuals in the workforce is not reported by BLS, and as of September 2017, BLS has no plans to begin collecting information on gender identity or sexual orientation in the annual Current Population

Survey.⁴ We believe that this is a critical oversight. A proper investigation of workforce demographics should be extended to individuals of all gender identities.

The present study uses data collected by Stack Overflow (SO) during their Annual Developer Survey (ADS). This data set addresses some of the issues aforementioned: it draws from a global population of software developers, allowing us to examine each country's gender representation over time. Moreover, it uses less limited categories for gender identification, allowing us to examine the representation of not-exclusively-male individuals (a more general category than women) among software developers.

Based on historical trends observed in the United States and summarized above, this study tests two different hypotheses. First, we examine whether there exists a positive marginal trend in the representation of not-exclusively-male individuals among software developers (indicating a greater representation of such individuals in this sector of the workforce). Historical data from the DOL reported above indicates that in years when women made up less than half of the general workforce, women's representation was consistently growing. Our first hypothesis investigates whether this effect (i.e., growth conditioned on less than 50% representation) generalizes to not-exclusively-male individuals in software development.

Second, we examine whether there is any correlation between each country's individual gender representation trajectory and their baseline gender proportion. This hypothesis is also based on historical data: the DOL reports that women's share of the civilian workforce has grown asymptotically toward a value of 50% in the U.S. Assuming such a pattern may be observed in other countries, we hypothesize that countries' random intercepts and random slopes will be negatively correlated such that a smaller baseline proportion of not-exclusively-male individuals is associated with a more positive longitudinal effect (i.e., a more rapid increase in the proportion of not-exclusively-male individuals). That is, we will observe on average a more rapid decrease in the proportion of males given that males constitute a larger proportion of a country's software developers and less change in the proportion of males over time given that males constitute about half of a country's software developers.

Throughout the paper we will go over our data-cleaning and manipulation, model, results, inferences, and finally limitations.

2 Methodology

2.1 Survey design

The ADS is a cross-sectional survey that has been conducted annually since 2011. Participation in ADS is voluntary and uncompensated. All interested individuals are able to access and participate in the ADS on the SO website (stackoverflow.com) provided that they have an internet connection and a capable web browser.

The ADS is conducted in the first quarter of every year, and it is open for volunteer participation for approximately one month. The publicly available information regarding the exact opening and closing dates of the ADS may be summarized in the following table for the years 2014 to 2018.

Year	Open date	Close date	Open duration*
2018	January 8	January 28	20 days
2017	January 12	February 6	25 days
2016	Unknown date in January	Unknown date in January	Unknown
2015	Early February	Mid February	Two weeks
2014	Unknown date between January 1 and February 19	Unknown date between January 1 and February 19	Unknown
*During this time, the ADS was available at all times of day.			

Table 2.1.1. Summary of ADS availability during the years 2014 through 2018.

The 2018 ADS consisted of at most 128 questions. Not all respondents were presented with every question: certain questions were not applicable to respondents on the basis of their responses to prior questions. For example, the question “Which of the following best describe your reasons for [participating in an online coding competition or hackathon]?” was only presented to respondents who indicated in a preceding question that they had participated in a hackathon. Most questions on the survey were optional and did not require a response, and this was true for all surveys analyzed in this study. There was some variation in the exact set of questions presented in the ADS across years, and, due to a lack of publicly available documentation, we cannot say whether the ADS redacted non-applicable questions in the years 2014 through 2017 as was done in 2018.

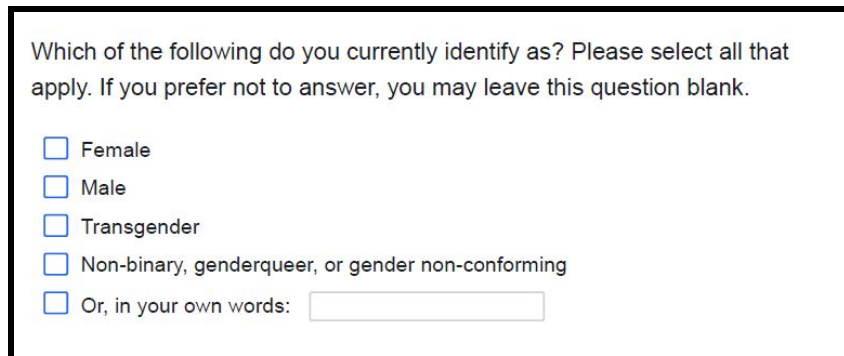
Respondents were allowed to complete the ADS at their own pace: no explicit time constraint was imposed. Thus, it was possible that some respondents completed the ADS over the course of several hours or several days. SO reported that the median time to completion of the 2018 survey was 29.4 minutes. Among “qualified” respondents (see Section 2.2), the median time to completion was 25.8 minutes for the 2018 ADS. Data on each respondent’s time to completion was not publicly available, and no median completion times were reported for years prior to 2018. We may reasonably assume that surveys took about 30 minutes to complete in all years.

In 2017, the ADS used a randomized block design such that blocks consisted of topically related questions, and each respondent was presented with a random sequence of blocks. This was done to reduce ordering effects; however, specific questions regarding respondents’ experiences with SO were always presented last. For lack of documentation, we cannot say whether the ADS used

a similar randomized blocking design in other years, but we may reasonably assume that they did.

In all years, the ADS was presented in English only. This admits a selection bias toward individuals who are English-speakers and who prefer to take surveys in English.

A respondent's level of *gender* was coded categorically. In 2017 and 2018, a respondent's recorded *gender* was based on a "check-all-that-apply" multiple choice question. An example of how this question appeared is provided below.



Which of the following do you currently identify as? Please select all that apply. If you prefer not to answer, you may leave this question blank.

- Female
- Male
- Transgender
- Non-binary, genderqueer, or gender non-conforming
- Or, in your own words:

Figure 2.1.1. An example of how the question of gender appeared to respondents completing the ADS.

Respondents' answers were recorded as a string (taken directly from the available responses above) corresponding to their selected gender identity. The gender identity of respondents who selected more than one identity was recorded as the concatenation of such strings.

In years prior to 2017, the question for gender identity accepted only a single response. The responses available during these years were "man", "woman", "other", and "prefer not to disclose". Respondents' answers were recorded as the single string selected from this set.

2.2 Analytical data set

SO reported that the 2018 ADS collected 101,592 "qualified" responses from software developers. The total number of responses ("qualified" or otherwise) collected in 2018 is not publicly available. So, it is not clear what percentage of all 2018 ADS responses were labeled as "qualified". According to SO, the "qualified" categorization was based on survey completion and time spent on the survey, but the exact criteria used for exclusion from the 2018 ADS are not publicly available. It is known that survey completion was not a criterion for "qualified" status since only 67,441 (66.4%) of "qualified" responses completed the entire survey. Notably, the open source data for the 2018 ADS consists of 98,855 responses, not 101,592. The source of this discrepancy is not clear.

The 2017 ADS collected data on a total of 64,227 software developers. Of these responses, 51,392 (80%) were deemed “usable”. According to SO, the “usable” designation applies to all completed surveys and to surveys labeled “partial complete”. The exact criteria used to determine whether a survey response was “partial complete” are unknown. The open source data for the 2017 ADS consists only of the 51,392 “usable” responses. The following table summarizes the amount for data available.

Inclusion-exclusion criteria for the ADS in years prior to 2017 were not publicly available. Thus, we must assume that the response numbers for those years reported in the table below represent all responses to the ADS in that year that SO deemed suitable for inclusion in the study overall. We may reasonably assume that inclusion criteria in these years were similar to the “qualified” and “usable” designations used in 2018 and 2017, respectively.

Year	Count of raw responses	Count of “qualified” or “useable” responses	Count of Responses in Open Data	Count of Countries Cited	Count of Countries in Open Data	Median response time (min)
2018	≥ 101,592	101,592	98,855	183	182	25.8
2017	≥ 64,227	51,392	51,392	213	200	26
2016	≥ 56,033	≥ 56,033	56,030	173	175	NA
2015	≥ 26,086	≥ 26,086	26,086	157	156	NA
2014	≥ 7,500	≥ 7,500	7,643	96	96	NA

Table 2.2.1. Summary of the amount of data available per year.

The variables most relevant to the present analysis are *year of response* and *gender* of the respondent. *Year of response* is not explicitly coded in each annual ADS data set, and it is not derived from any of the survey responses. *Year of response* is obtained trivially from the year of the data set that contains each response. Thus, the *year of response* for all data points in the 2018 ADS open source raw data is “2018”. A similar statement is true for data points in the open source raw data for prior years.

The other variable of interest is the *gender* of the respondent. Given that SO used a randomized blocking design to determine the sequence of questions presented to each respondent in 2017 (and possibly 2018), it is possible that this question was answered earlier in the survey for some respondents compared to others. The exact order of questions presented to each respondent was not recorded in any year.

The process for recoding gender of respondent from the values recorded in the manner described in Section 2.1 was based on the following definitions. A *male* is any person who identifies

themselves on the ADS as exclusively male; since respondents may provide more than one gender identity on the ADS, we specify that a male is a respondent who indicated male as their only gender. An *other* is any person who identifies themselves on the ADS as any gender other than exclusively male (including combinations of genders that do or do not include male). The sample space for gender $\Omega = \{\text{male, female, non-binary, transgender, gender non-conforming, genderqueer, other}\}$.

2.2.1 Data Cleaning

Since individual respondents cannot be tracked over time in our data set, our modeling procedures (described in Section 2.3) treat countries as individuals. In order to narrow down on a final subset of our original data, we needed to do two things: isolate full-time respondents and narrow down on a list of countries that were consistent across the all years of interest.

Isolating full-time respondents was simple for 2015-2018 since employment status was a question on all those surveys. Although 2014 did not have such a question, we were able to create a derived variable that estimated the amount of hours per week each respondent worked. There was a series of questions that asked each respondent to allocate a range of hours for which they worked in several categories (e.g. developing new features, fixing bugs, meetings, etc.). After taking the sum of the maximum of each range, we decided that all respondents who noted hours larger than 30 were considered full-time software developers. After we narrowed down on full-time employees, our number of respondents decreased an average of 29% from the original dataset.

Unionizing countries across all years gave us 219 uniquely named countries. However, after scanning through the data from ADS for years 2014 through 2018, we found that not all country names were consistent across the years. In fact, there were 3 countries that were misspelled at least once ('Azerbaijan' was included in 2017, but is spelled without a 'd') and 19 countries that had at least one equivalent name (e.g. 'Moldova' included in 2014, 2015, and 2016 is also called 'Republic of Moldova', which is included in 2018). After we cleaned that part of the data, we set certain thresholds that led us to our final dataset. In longitudinal analysis, we can only use countries that are represented in at least 3 out of the 5 years in order to create an estimated trajectory for each country. In addition, number of respondents for each year and each country needed to be at least greater than or equal to 30 for at least 3 years in order to be included in the dataset. Thirty was chosen as a threshold in light of Central Limit Theorem. Our final sample size was 63 countries.

The above sample size is based on complete cases only (respondents missing gender were excluded from our study). In our Statistical Analysis Plan, we anticipated performing imputation on missing gender values. After some consideration, we determined that such imputation would not be viable due to a large degree of missingness in other covariates. Moreover, the set of countries included in our study by the criteria described above provide a large and representative sample of advanced economies. It is unlikely that this sample would be made substantially larger

by imputing gender values (and doing so might severely bias the gender proportions for poorly represented countries). A missing values analysis is reported in Section I of the appendix as an assessment of the degree to which missing gender may be predicted by observed covariates, and any limitations on the conclusions of our study due to missingness are discussed in Section 4.2.

2.3 Modeling procedures

Our primary analysis involved a single general linear mixed model (GLMM). The GLMM approach allowed us to model correlated binomial outcomes over time. Given n countries in our sample, and m years for which we had data on each country, the model statement for the primary analysis was as follows:

$$Y_{it} = \mu_{it} + \varepsilon_{it} = \text{expit}(\beta_0 + \beta_{year}(t - t_{i0}) + \beta_{base}p_{i0} + b_{0,i} + b_{1,i}(t - t_{i0}) + \varepsilon_{it})$$

$$\text{for } i = 1, \dots, n, \quad t = 0, \dots, m - 1 ;$$

$$\vec{b}_i \sim N(0, D) \text{ and } \vec{\varepsilon}_i \sim N(0, R_i).$$

In the model statement above, μ_{ij} is the proportion of *male* individuals for the i^{th} country at year j . The constant β_0 is the marginal intercept, the coefficient β_{year} is the marginal longitudinal effect, and the coefficient β_{base} is the effect of baseline proportion. The vector \vec{b}_i is a vector of random effects such that $b_{0,i}$ is the random intercept for the i^{th} country and $b_{1,i}$ is the random slope for the i^{th} country. The vector \vec{b}_i is distributed multivariate normal with 2×2 variance-covariance matrix D to be estimated. The variance-covariance R_i matrix of $\vec{\varepsilon}_i$ will also be estimated, and \vec{b}_i and $\vec{\varepsilon}_i$ are assumed independent. The residuals $\vec{\varepsilon}_i$ will be parameterized according to a banded autoregressive correlation structure. That is,

$$\text{Cov}[\varepsilon_{ij}, \varepsilon_{ik}] = \sigma^2 \rho^{|j-k|}$$

where σ^2 gives the within-country variance, and ρ gives the correlation between two consecutive observations on a particular country.

This model assumes that there is a global trajectory in the change in proportion of male individuals in full-time software development positions over time, and it assumes that every country randomly deviates from this global trajectory. This allowed us to assess whether there was any change in the proportion of males over time on average across countries.

Based on reports from the DOL spanning several decades, women's share of the civilian workforce has grown asymptotically toward a value of 50% in the U.S. This suggests that the rate of change in the proportion of males in the workforce at any given time depends on the proportion at that time. We posit that such a pattern can be observed in the growing share of women in the workforce in other countries. For this reason, we claim that our model, which

allows for correlated random effects, is the correct model for the data and provides the most precise estimates.

Moreover, it is likely that any two measurements taken closely in time on one country will be more highly correlated than two measurements taken farther apart. This is because the present workforce is directly related to the workforce of the immediate past. This relationship deteriorates with increasing temporal distance, so an autoregressive structure on the variance of errors is appropriate and maximizes precision.

With regard to our scientific objectives, we were interested primarily in estimating the marginal longitudinal effect β_1 , each country's deviation from the marginal effect $b_{1,i}$, and the covariance between the random effect and random intercept $D_{1,2}$. Since we were less interested in estimating the within-country variance or the correlation between consecutive observations on a country over time, we chose not to generate a variogram as originally anticipated by our Statistical Analysis Plan. It is thus possible that despite the justification given in the previous paragraph, our covariance structure was misspecified. Nonetheless, our estimates should be consistent (if not maximally efficient) regardless of a potential misspecification of the covariance structure.

As is common in the GLMM context, our model was fit using MCMC as implemented within the R2jags library available for R. Each observation for each country was weighted according to the number of individual responses recorded from that country during that year of the ADS. Prior distributions on the model are given by the following vague proper priors:

$$\begin{aligned}
 \beta_0 &\sim N(a, b) \\
 \beta_{year} &\sim N(a, b) \\
 \beta_{base} &\sim N(a, b) \\
 a &\sim N(0, 1000) \\
 b^{-1} &\sim \Gamma(0.001, 0.001) \\
 D &\sim InvWish(4, -1 + 2I_{2 \times 2}) \\
 \sigma^{-1} &\sim \Gamma(0.001, 0.001) \\
 \rho &\sim Unif(-1, 1)
 \end{aligned}$$

All of the above prior and hyperprior specifications are commonly used in Bayesian statistical practice to minimize the influence of the priors; this, in turn, minimizes bias in our estimate.

We performed MCMC in two passes. On the first pass, we ran 4 chains for 10,000 iterations with an initial burn-in of 0. During this pass, we recorded the samples of $\hat{\beta}_0$ and $\hat{\beta}_1$ on each iteration. The resultant MCMC samples were used to compute a scale reduction factor \hat{R} from the samples of $\hat{\beta}_0$ and $\hat{\beta}_1$. The value of \hat{R} is a summary comparison of within-chain variance to between-chain variance, and it is standard to run MCMC until \hat{R} stays below 1.1 before using subsequent MCMC samples for parameter estimation. We generated a plot of \hat{R} across the 10,000 iterations to identify the number of iterations necessary to meet this criterion. At the same

time, we generated an autocorrelation plot from the samples of $\hat{\beta}_0$ and $\hat{\beta}_1$ to check that the absolute value of the autocorrelation was less than 0.1 at lags greater than 0. If the above two criteria were not met within the first 10,000 iterations, we resumed the MCMC process in increments of 1,000 iterations until they were met.

The number of iterations necessary to meet both of the above criteria was used as the burn-in for the second pass MCMC. Ultimately, we found that the above criteria were met using 50,000 iterations, 4 chains, a thinning rate of 100, and a burn-in of 10,000. The final parameter estimates were based on the samples collected from the second pass MCMC. During this pass, we also recorded the samples of each country's random intercept estimates and random slope estimates on each iteration.

Since we chose not to perform imputation on the data (for reasons described in Section 2.2), we did not conduct any sensitivity analyses to check the influence of imputed data. Instead, we proceeded with complete cases only. To examine the sensitivity of our estimates to our priors, we fit the model specified above with different hyperparameters for the marginal effects. We considered the following alternate parameterizations of the hyperpriors (every combination of one choice from each set):

$$\{ a \sim N(0, 1); a \sim N(0, 10); a \sim N(0, 100); a \sim N(0, 1000) \}$$

$$\{ b^{-1} \sim \Gamma(1, 1); b^{-1} \sim \Gamma(0.1, 0.1); b^{-1} \sim \Gamma(0.01, 0.01); b^{-1} \sim \Gamma(0.001, 0.001) \}$$

We found little evidence that our parameter estimates were sensitive to our prior specifications. Thus, the results reported below come from fitting the exact model specified previously.

3 Results

The MCMC algorithm seemed to converge well. We run 4 chains, each with 50000 iterations and discard the first 10000 samples for each chain. For each chain we use every 100th sample. Autocorrelation plots showed little long-lag autocorrelation between chains.

Table 3.1 shows the marginal effects on the odds of males. Overall, the proportion of males appear to decrease with time. The odds of males was estimated to decrease by a mean of 10.9% each year, with chance over 95% to take values between 4.3% and 17.3%, and the estimated probability for the actual effect to be non-negative is less than 0.001. Note that baseline proportion of males have a strong positive effect on proportion of males: each percent of positive difference in baseline proportion of males could result in around 3.8 percent of positive difference in odds of males.

Predictor	Mean	sd	95% CI	P(>0)
Baseline	3.774	0.725	(2.295, 5.124)	>0.999
Year	-0.115	0.037	(-0.190, -0.044)	<0.001
(Intercept)	-0.584	0.696	(-1.832, 0.825)	-

Table 3.1. Summary of the Bayesian GLMM fixed effects.

Countries	Counts	Mean	2.5 th Percentile	97.5 th Percentile
United States	33053	0.871	0.799	0.928
India	11897	0.933	0.892	0.965
United Kingdom	11016	0.91	0.856	0.95
Germany	8347	0.905	0.85	0.949
Canada	4897	0.868	0.793	0.924
Dominican Republic	136	0.952	0.903	0.98
Uruguay	156	0.939	0.888	0.974
Taiwan	164	0.877	0.778	0.941
Thailand	173	0.922	0.856	0.964
Estonia	190	0.891	0.778	0.952

Table 3.2. Prediction of Mean Proportion of Males as of 2019 for 5 most observed (largest counts of respondents over 2014-2018) and 5 least observed (smallest counts of respondents over 2014-2018) countries.

For individual countries, the general trend for proportion of males was increasing with a few exceptions, and there seemed to be a lot of variability between countries. 57 out of 63 have mean proportion of males decreasing over time(negative combined slopes), with a rate varying from 4.6% to 28.2% a year with respect to odds of males. For countries with increasing proportion of males, the highest rate of increase is at 9.3% per year. It is worth noticing that countries vary a lot in their rate of change in proportion of males. The variance of rates is 0.0425, which is high relative to the scale of rate of change.

Countries	Respondents	Years Observed	Observed	Predicted	2.5th Percentile	97.5th Percentile
United States	20377	2014-2018	0.898154	0.898123	0.834975	0.940895
United Kingdom	7409	2014-2018	0.927086	0.946306	0.914366	0.969033
India	6686	2014-2018	0.924007	0.915971	0.868485	0.951578
Germany	5232	2014-2018	0.94382	0.913713	0.867775	0.950496
Canada	3047	2014-2018	0.895676	0.877239	0.817523	0.926368
Dominican Republic	80	2016-2018	1	0.957685	0.910757	0.984669
Thailand	87	2016-2018	0.906977	0.957529	0.905106	0.985287
Taiwan	88	2016-2018	0.855263	0.941405	0.871568	0.977388
Chile	99	2016-2018	0.918367	0.927534	0.848399	0.973146
Estonia	99	2016-2018	0.956044	0.94518	0.884779	0.979634

Table 3.3. Part of Prediction of Proportion of Males 2018 using data in 2014-2017. Countries shown are the 5 most observed and 5 least observed countries(based on integrated number of individual observations).

Posterior Median (95% PI) of Country-Specific Intercepts

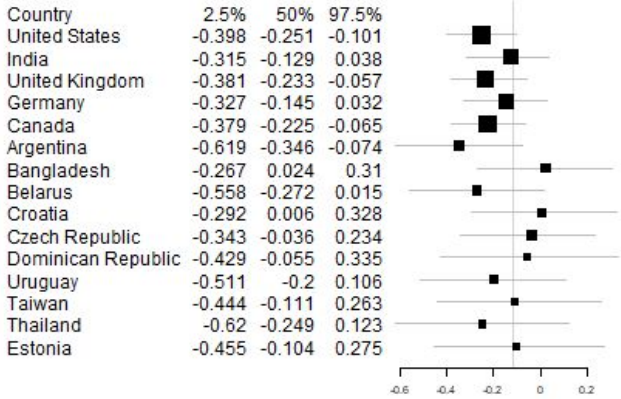


Figure 3.1. Forest-plot of 5 most observed countries, 5 least observed countries and 5 randomly sampled countries(each observed 4 times out of 5)

Figure 3.1 shows a forest-plot of country-specific slopes for 15 countries, including 5 most observed countries (based on the integrated counts of observations over the 5 years), 5 least observed countries and 5 randomly sampled countries, each satisfying the constraint of being observed 4 years out of 5; number of observations is distinguished by size of diamond marks. Of the 15 countries, only 8 countries showed significant negative change in proportion of males over time with 95% credible intervals on the individual slopes completely in negative range : Argentina, US, UK, Canada, Australia, China, Austria, and Romania. All other countries had inconclusive 95% credible intervals of their slopes, containing negative and positive values.

4 Discussion

4.1 Marginal trend in gender representation

As stated in Table 3.1, we find strong significant evidence of a negative global trend in the proportion of male software developers across the years 2014 through 2018. We estimate that the relative odds of being a male software developer in the countries represented by our sample decreases by 10.9% each year, on average, with a 95% credible interval between 4.3% and 17.3%. This result supports our hypothesis that there exists a positive marginal trend in the representation of not-exclusively-male individuals among software developers, reflecting trends observed in historical data regarding the representation of women in the general civilian workforce in the United States.

Moreover, we estimate that $D_{1,2}$ (the covariance between the random slopes and random intercepts) is -0.0435 with 95% credible interval (-0.0753, -0.0255). This provides significant evidence that countries' random intercepts and random slopes are negatively correlated. This finding supports our hypothesis that a smaller baseline proportion of not-exclusively-male individuals is associated with a more positive longitudinal effect (i.e., a more rapid increase in the proportion of not-exclusively-male individuals). From this result, we might predict that as proportions grow toward 50% male, 50% not-exclusively-male, the change in gender proportions will tend toward zero. However, no country in our study had male/not-exclusively-male ratios near the 50/50 mark, so it is possible that the change in gender representation will tend toward zero at a ratio other than 50/50. Indeed, it may be argued that the DOL data cited in our introduction present such a scenario. The source of such a gender employment gap is worth further investigation, but cannot be addressed by these data.

For the Bayesian GLMM model, it is worth noticing that the accuracy of results with respect to each country seemed to be dependent on the number of observations available. In Figure 3.1, Countries observed in more years or with more observations tend to have lower variance and smaller credible intervals. This also holds in accuracy of prediction. Table 3.3 shows part of the prediction test results for proportion of males in 2018 using data from 2014 to 2017. The predicted mean proportion of males is very close to the true value for the US, which has over

20,000 individual observations. For countries less observed like Dominican Republic (80 observations in 2 years), the observed proportion of males is not even in the 95% credible interval. The mean prediction error for countries observed for only 2 years is over 7 times that of countries observed every year.

4.2 Gender missing at random

Missing gender in 2014 through 2016 is ignorable. Missing gender in 2017 and 2018 is associated with missingness in other demographic values. This may be interpreted as evidence that missingness is a function of question order (since demographic questions appear to have been asked together); however, it is also possible that the association may be explained by the latent confidentiality of respondents (a tendency to not provide personally identifiable information on a survey). Still, it is unlikely that these two possibilities alone might explain the pattern of missing gender entirely.

Missing gender percentages in 2017 are strongly correlated within-country with missing gender percentages in 2018. This suggests that the unknown source of missingness affects each country differently, in a consistent way over time (or at least over the last two years). For example, perhaps there is a cultural propensity to avoid answering such questions. Regardless of the source of the correlation between 2017 and 2018, this correlation is not what we would expect if the primary source of missingness was question order.

We see that missing gender is associated with lower salary. However, differences in mean country salary do not explain differences in missing gender percentages across countries. Missing gender is also associated with level of education (in both 2017 and 2018) and undergraduate major (in 2018 only).

We must conclude that gender data is not missing completely at random (MCAR). Missing gender is associated with lower salary, level of education, and undergraduate major. Moreover, percent respondents missing gender in 2018 is associated within-country with the corresponding percentage in 2017. Missing data is clearly related to some of our observed data, so we can only conclude that we have a missing at random (MAR) scenario, at best.

5 Conclusion

Although our results seem to imply that some countries are more progressive at eliminating gender bias in the tech workforce, we need to remind ourselves that there are unignorable biases in the dataset. Most surveys are notoriously biased in more than one way. Our survey, though rich with information, is only available at a certain time of the year, advertised on particular platforms, and majorly popular among the StackOverflow community. This means that our results are, at best, representative of the StackOverflow full-time tech workers who are not too busy and actually interested in completing a 30 minute online survey during the beginning of the

new year. It is also reasonable to assume that not all respondents answered truthfully to all the questions. If voluntary response bias and convenience sampling weren't enough, there is a large amount of missing data. We proved that our missingness is at best MAR, and usually imputation is the best solution. Though, we continue with complete case because there is too much missingness in other covariates that would otherwise allow us to impute via regression methods. In the future, the best medicine against missingness is constructing better surveys. We believe StackOverflow has made an effort each year to improve their sampling techniques with more advertisements on a wider variety of platforms. Overall, we believe our attempts at doing longitudinal analysis on such a dataset will help convince others of the need for better data on this topic. We hope that our paper will influence a further investigation into the tech workforce gender bias problem, which definitely includes figuring out a way to collect more representative data and prevent missing values.

6 Acknowledgements

We would like to express our gratitude to Professor Hal Stern for reviewing our modeling procedures and sharing his expert advice. We express our thanks also to Professor Thomas Lin and the students of Stats 275 in Fall Quarter 2018 for their helpful commentary.

7 References

[1] Women's Bureau, United States Department of Labor. (n.d.). Civilian labor force by sex, 1948-2016 annual averages. Retrieved October 30, 2018, from https://www.dol.gov/wb/stats/NEWSTATS/facts/women_lf.htm#one.

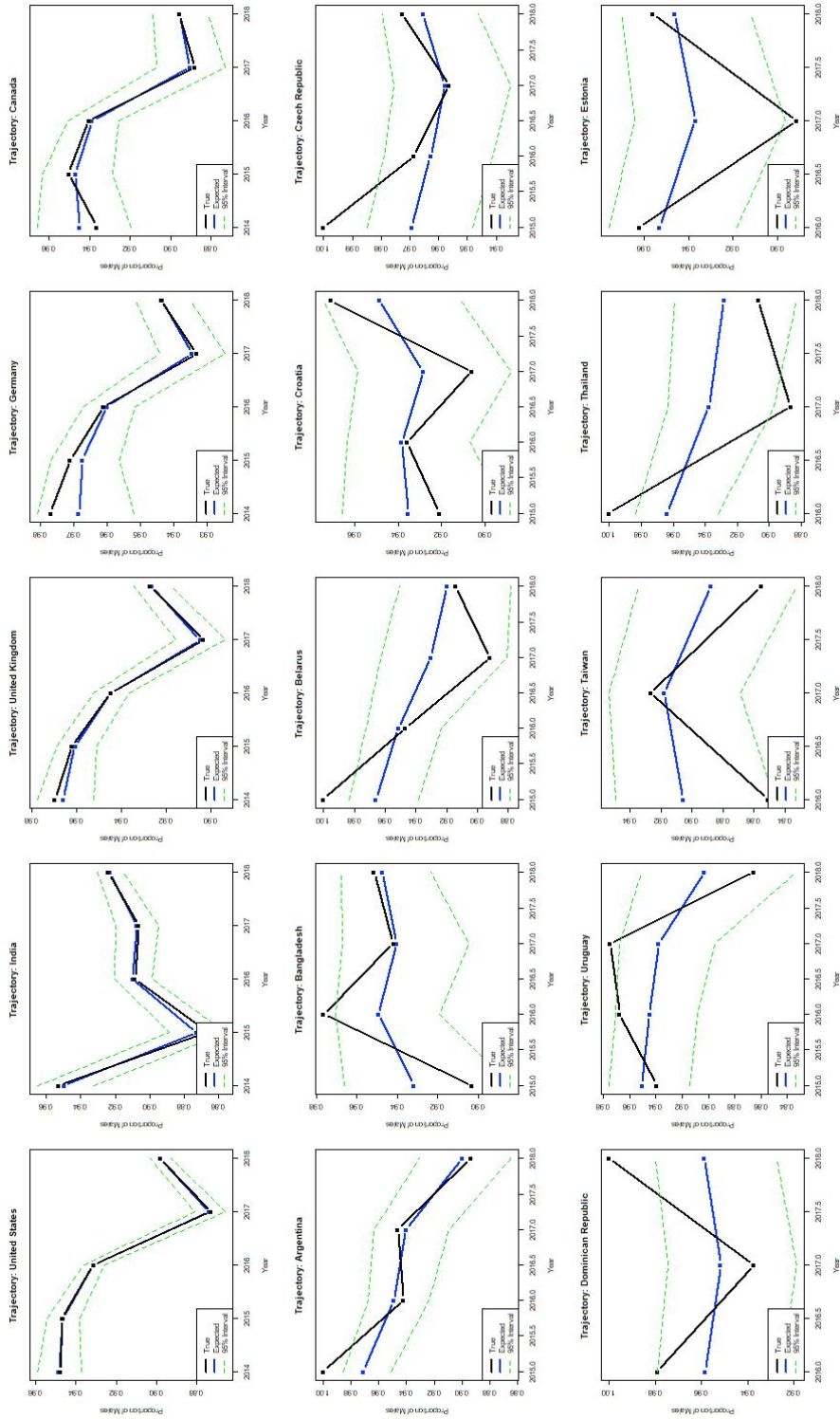
[2] Bureau of Labor Statistics. (n.d.). Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity [for 2017]. Retrieved October 30, 2018, from <https://www.bls.gov/cps/cpsaat11.htm>.

[3] Bureau of Labor Statistics. (n.d.). Employed persons by detailed occupation, sex, and age [for 2015/2016]. Retrieved October 30, 2018, from <https://www.bls.gov/cps/aa2016/cpsaat09.htm>.

[4] Bureau of Labor Statistics. (n.d.). Assessing the feasibility of asking about sexual orientation and gender identity in the Current Population Survey. Retrieved October 30, 2018, from https://www.bls.gov/osmr/pdf/cps_sogi_executive_summary.pdf.

Appendix

I. Trajectory Plots



II. Primary information: Missing Values Analysis

For each country represented by our sample, less than 10% of respondents were missing a value for gender in the years 2014, 2015, and 2016. This amount of missingness was deemed ignorable, and so no further analysis of missing data from those years was performed.

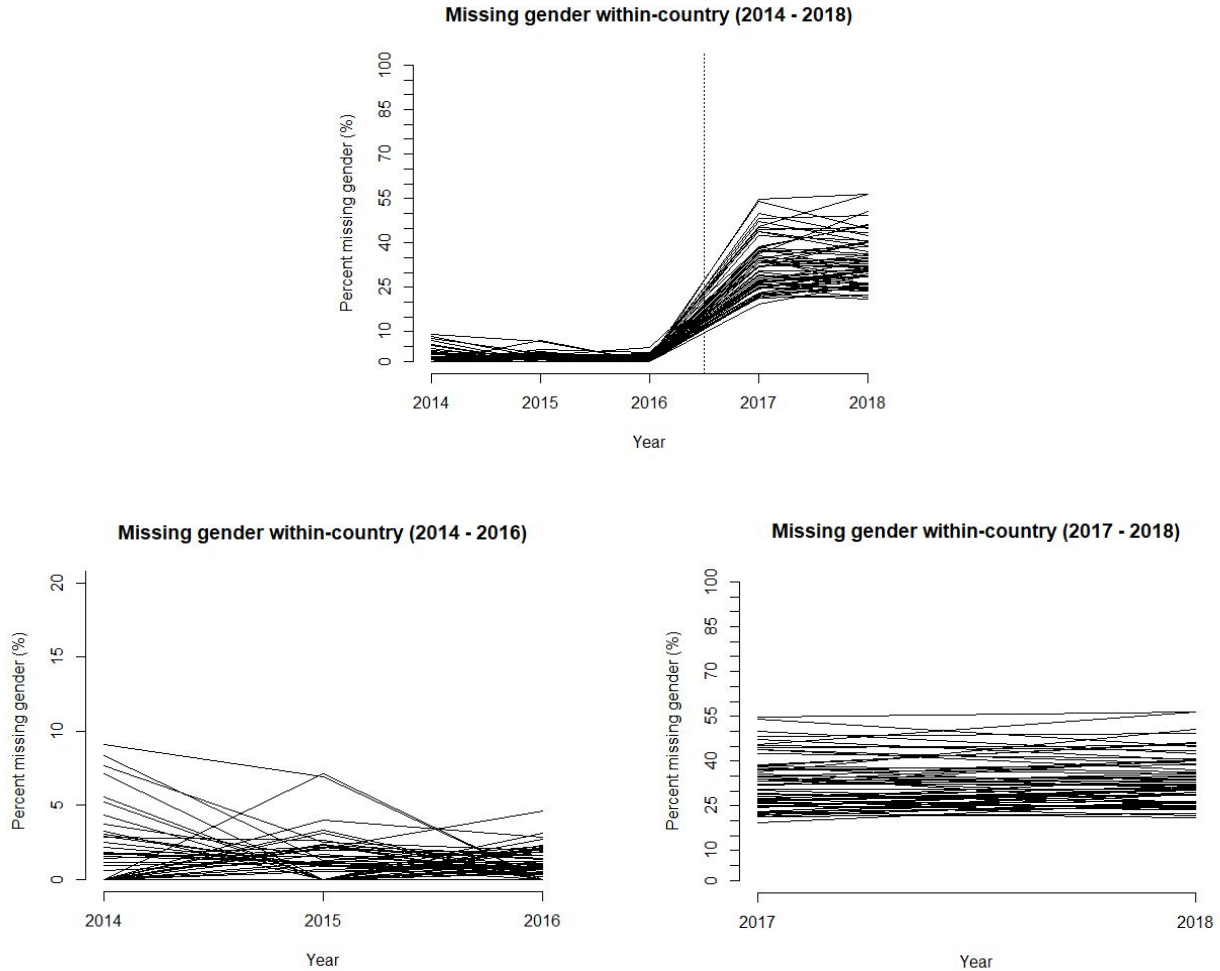


Figure A. Spaghetti plots that show each country's percent respondents missing gender for all years considered (top), 2014 through 2016 only (bottom left), and 2017 and 2018 (bottom right)). Each line represents a single country. The dotted line in the top graph indicates the boundary between the data of the bottom two graphs.

In 2017 and 2018, more than 10% of respondents were missing gender for each country, with an average of about 33%. The following table summarizes the distribution of percent respondents missing gender within-countries in 2017 and 2018.

	2017	2018
N	63	62
Minimum	19.30	20.93
Q1	26.09	26.17
Median	31.88	31.25
Mean	32.35	33.43
Q3	37.08	39.04
Maximum	54.67	56.58

Table A. Summary of the distribution of percent respondents missing gender within-countries in 2017 and 2018

Using the 1.5-IQR rule to identify outliers in the 2017 and 2018 data, we found two outliers in 2017 (Vietnam [54.67% missing] and South Korea [53.98% missing]) and no outliers in 2018.

Percent respondents missing gender within-countries in 2018 was correlated with the corresponding percentages in 2017.

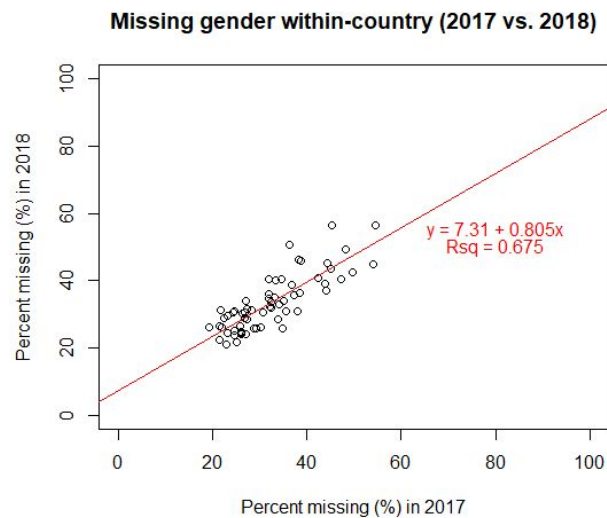


Figure B. Percent respondents missing gender within-countries in 2018 regressed on corresponding percentages in 2017. Estimated line equation and R-squared of the line are reported in red.

Missing gender appeared to be associated with a lower salary. Note that this particular association was based on subjective assessment. Given that we are in a large sample setting, standard hypothesis testing methods would have yielded significant results for even very small differences. Moreover, the 2018 Annual Developer Survey allowed respondents to report salary in units other than USD. Thus the mean and standard deviations reported above for 2018 were based on conversion from reported value and currency. The following table reports the mean salary for individuals based on the missingness of their gender value.

Gender	2017 Salary (USD)		2018 Salary (USD)*	
	Mean	SD	Mean	SD
Not missing	59805.12	39388.84	60115.61	41641.88
Missing	50914.77	40783.45	48376.68	41296

Table B. Summary of the mean and standard deviation of respondents' salary in each year (2017 and 2018), grouped by gender missingness in the corresponding years.

A series of chi-squared tests provided evidence that missing gender was associated with undergraduate major (STEM versus non-STEM) in both 2017 and 2018. It was associated with level of education in 2018, but not in 2017. Detailed tables summarizing the data may be found in Section II of the appendix.

Missing gender was also associated with missingness in other covariates. Missing gender was very strongly associated with missing salary and missing undergraduate major in both 2017 and 2018. It was associated with missing age in 2018 (age was not reported in 2017, so we cannot say whether or not an association exists for that year). Level of education was not missing any values in 2017, but in 2018, missing level of education was associated with missing gender. Detailed tables summarizing the frequency of missing covariates may be found in Section I of the appendix. The following table summarizes tests of independence between gender and levels of each covariate, as well as missingness in each covariate.

Covariate	2017			2018		
	df	Chi-sq	P	df	Chi-sq	P
Age	-	-	-	6	4.87	0.56
Level of education	5	60.1	<0.001	6	232.61	<0.001
Undergraduate major	1	0.39	0.53	1	50.25	<0.001
Missing salary	1	3534.5	<0.001	1	28139	<0.001

Missing age	-	-	-	1	61041	<0.001
Missing level of education	1	5963.7	<0.001	1	55.8	<0.001
Missing undergraduate major	1	36.5	<0.001	1	1347.5	<0.001

Table C. Summary of the chi-squared tests for independence between missing gender and several other covariates. Degrees of freedom are based on the rows and columns of the corresponding frequency table.

III. Supplementary information: missing values analysis

Percent missing gender by age group (2018)

Under 18 years old	18 - 24 years old	25 - 34 years old	35 - 44 years old	45 - 54 years old	55 - 64 years old	65 years or older
3.7	1.78	1.7	1.56	1.73	1.51	4.35

Frequency of missing gender by age group (2018)

	Not missing gender	Missing gender
Under 18	52	2
18 – 24	7980	145
25 – 34	25591	442
35 – 44	9168	145
45 – 54	2379	42
55 – 64	589	9
Older than 65	44	2

Percent missing gender by level of education (2017 and 2018)

	2017	2018
No college	31.78	36.50
Some college, no degree	26.33	26.99
Professional degree	39.69	39.77

Associate degree	Not represented in 2017	31.98
Bachelor's degree	29.33	31.61
Master's degree	29.27	31.40
Doctoral degree	26.00	21.07

Frequency of missing gender by level of education (2017)

	Not missing gender	Missing gender
No college	1638	763
Some college, no degree	3671	1312
Professional degree	307	202
Bachelor's degree	11686	4850
Master's degree	6322	2616
Doctoral degree	797	280

Frequency of missing gender by level of education (2018)

	Not missing gender	Missing gender
No college	2375	1365
Some college, no degree	5122	1893
Associate degree	1393	655
Professional degree	618	408
Bachelor's degree	23053	10657
Master's degree	11868	5433
Doctoral degree	1435	383

Percent missing gender by undergraduate major (2017 and 2018)

	2017	2018
STEM	28.89	29.33

Non-STEM	28.26	23.66
----------	-------	-------

Frequency of missing gender by undergraduate major (2017)

	Not missing gender	Missing gender
STEM	21129	8586
Non-STEM	1640	646

Frequency of missing gender by undergraduate major (2018)

	Not missing gender	Missing gender
STEM	39544	16412
Non-STEM	2623	813

Percent missing gender by missingness of salary

	2017	2018
Not missing salary	8.81	6.36
Missing salary	39.58	67.14

Frequency of missing gender and missing salary

	2017		2018	
	Not missing gender	Missing gender	Not missing gender	Missing gender
Not missing salary	10603	1025	37334	2537
Missing salary	13959	9145	9159	18710

Percent missing gender by missingness of age

	2018
Not missing age	1.69
Missing age	96.74

Frequency of missing gender and missing salary

	2018
--	------

	Not missing gender	Missing gender
Not missing age	45803	787
Missing age	690	20460

Percent missing gender by missingness of level of education

	2017	2018
Not missing level of education	29.28	31.20
Missing level of education	0.00	41.87

Frequency of missing gender and missing level of education

	2017		2018	
	Not missing gender	Missing gender	Not missing gender	Missing gender
Not missing level of education	24562	10170	45864	20794
Missing level of education	0	0	629	453

Percent missing gender by missingness of undergraduate major

	2017	2018
Not missing major	28.85	28.99
Missing major	34.35	49.38

Frequency of missing gender and missing major

	2017		2018	
	Not missing gender	Missing gender	Not missing gender	Missing gender
Not missing major	22769	9232	42493	17345
Missing major	1793	938	4000	3902