**Cleaning and exploring the salary data**

In this post, I use the most recent year’s salary data to explore how salary is affected by the **gender** of the respondent. You can download the public version of the [2019 results yourself](https://insights.stackoverflow.com/survey/); that is the version of the dataset I’ll use in this analysis. In this first step, I open up the CSV file and filter to only respondents in the US who are employed full-time as individual contributors in industry (i.e. not academics and not upper level managers). It would be interesting to extend this analysis to other countries, and to understand how salaries in academia are impacted by gender, but I want to define this question narrowly to set this up for success as much as possible.

library(tidyverse)

filtered\_gender <- c("Man", "Woman", "Non-binary")

survey\_results\_raw <- read\_csv("survey\_results\_public.csv") %>%

filter(

Country == "United States",

Employment == "Employed full-time",

ConvertedComp > 3e4,

ConvertedComp < 2e6

)

## identify non-ICs, to remove

managers\_ctos <- survey\_results\_raw %>%

filter(str\_detect(DevType, "Engineering manager|Product manager|Senior executive/VP"))

## identify academics, to remove

academics <- survey\_results\_raw %>%

filter(str\_detect(DevType, "Academic researcher|Scientist|Educator"))

survey\_results <- survey\_results\_raw %>%

anti\_join(managers\_ctos) %>%

anti\_join(academics) %>%

transmute(Respondent,

EdLevel = fct\_collapse(EdLevel,

`Less than bachelor's` = c(

"I never completed any formal education",

"Primary/elementary school",

"Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)",

"Some college/university study without earning a degree",

"Associate degree"

),

`Bachelor's degree` = "Bachelor’s degree (BA, BS, B.Eng., etc.)",

`Graduate degree` = c(

"Other doctoral degree (Ph.D, Ed.D., etc.)",

"Master’s degree (MA, MS, M.Eng., MBA, etc.)",

"Professional degree (JD, MD, etc.)"

)

),

DevType,

OpenSourcer = fct\_collapse(OpenSourcer,

Never = "Never",

Sometimes = "Less than once per year",

Often = c(

"Less than once a month but more than once per year",

"Once a month or more often"

)

),

OpenSourcer = fct\_rev(OpenSourcer),

YearsCodePro = parse\_number(YearsCodePro),

Gender = case\_when(

str\_detect(Gender, "Non-binary") ~ "Non-binary",

TRUE ~ Gender

),

Dependents,

ConvertedComp

) %>%

filter(Gender %in% filtered\_gender)

survey\_results

## # A tibble: 10,546 x 8

## Respondent EdLevel DevType OpenSourcer YearsCodePro Gender Dependents

##

## 1 4 Bachel… Develo… Never 1 Man No

## 2 22 Less t… Data o… Sometimes 18 Man Yes

## 3 23 Bachel… Develo… Sometimes 1 Man No

## 4 32 Less t… Develo… Never 2 Man No

## 5 35 Bachel… Develo… Never 14 Woman No

## 6 39 Bachel… Databa… Sometimes 23 Man No

## 7 53 Less t… Develo… Often 3 Man No

## 8 60 Less t… Develo… Often 9 Man No

## 9 95 Bachel… Develo… Sometimes 8 Man Yes

## 10 105 Bachel… Design… Sometimes 2 Woman No

## # … with 10,536 more rows, and 1 more variable: ConvertedComp

You can read more about the [survey and analysis methodology](https://insights.stackoverflow.com/survey/2019#methodology) for the 2019 Stack Overflow Developer Survey. One step in that data analysis is how very high salaries were dealt with; the top approximately 2% of salaries inside and outside of the US were trimmed and replaced with threshold values. More men than folks of other genders reported salaries above this threshold in the US, but I removed these very high, thresholded salaries (along with low salaries that are unlikely to be true full-time salaries in the United States) for this analysis.

What does the distribution of salaries look like by gender? The gender identification question allowed respondents to select all that apply, and the question about transgender identity was separate from the question about gender identity, so that, for example, transgender women would be included under the category of women overall.

survey\_results %>%

ggplot(aes(ConvertedComp, fill = Gender, color = Gender)) +

geom\_density(alpha = 0.2, size = 1.5) +

scale\_x\_log10(labels = dollar\_format()) +

labs(

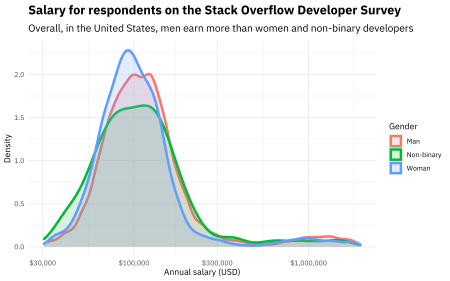
x = "Annual salary (USD)",

y = "Density",

title = "Salary for respondents on the Stack Overflow Developer Survey",

subtitle = "Overall, in the United States, men earn more than women and non-binary developers"

)



Salaries for men are shifted higher compared to the salaries for women while the distribution for non-binary respondents is broader. There are 175 people who identified as non-binary in this particular sample.

survey\_results %>%

group\_by(Gender) %>%

summarise(

Total = n(),

Salary = median(ConvertedComp)

) %>%

arrange(Salary) %>%

mutate(

Total = comma(Total),

Salary = dollar(Salary)

) %>%

kable(

align = "lrr",

col.names = c("Gender", "Total respondents", "Median annual salary for US respondents")

)

| **Gender** | **Total respondents** | **Median annual salary for US respondents** |
| --- | --- | --- |
| Woman | 1,127 | $98,000 |
| Non-binary | 175 | $100,000 |
| Man | 9,244 | $108,000 |

Why do we see this difference in salaries? Is it a meaningful difference? What is contributing to it? The situation is complicated, because we know that the [women who responded to this survey are less experienced](https://insights.stackoverflow.com/survey/2019#developer-profile-_-experience-and-gender) than the men who responded (along with non-binary developers, although less dramatically). Also, people of different genders [participate in different kinds of coding work at different levels](https://insights.stackoverflow.com/survey/2019#developer-profile-_-developer-role-and-gender), and these different kinds of work are compensated differently. Women are [less likely to say they participate in open source](https://insights.stackoverflow.com/survey/2019#developer-profile-_-contributing-to-open-source) as well. We have evidence from existing work that all of these factors influence salary, but I would like to know which ones may explain the observed differences in salaries and if there is any remaining difference after accounting for such characteristics. To make the modeling goal a little more explicit, **do we see evidence that gender affects salary in the United States, controlling for experience, type of coding work, dependents, and open source contributions?**

**Impact of developer role and experience**

Before we start building models, let’s do a bit more exploratory work.

filtered\_devtype <- c(

"Other", "Student",

"Marketing or sales professional"

)

survey\_results\_parsed <- survey\_results %>%

mutate(DevType = str\_split(DevType, pattern = ";")) %>%

unnest(DevType) %>%

mutate(

DevType = case\_when(

str\_detect(str\_to\_lower(DevType), "data scientist") ~ "Data scientist",

str\_detect(str\_to\_lower(DevType), "data or business") ~ "Data analyst",

str\_detect(str\_to\_lower(DevType), "desktop") ~ "Desktop",

str\_detect(str\_to\_lower(DevType), "embedded") ~ "Embedded",

str\_detect(str\_to\_lower(DevType), "devops") ~ "DevOps",

str\_detect(DevType, "Engineer, data") ~ "Data engineer",

str\_detect(str\_to\_lower(DevType), "site reliability") ~ "DevOps",

TRUE ~ DevType

),

DevType = str\_remove\_all(DevType, "Developer, "),

DevType = str\_to\_sentence(DevType),

DevType = str\_replace\_all(DevType, "Qa", "QA"),

DevType = str\_replace\_all(DevType, "Sre", "SRE"),

DevType = str\_replace\_all(DevType, "Devops", "DevOps")

) %>%

filter(

!DevType %in% filtered\_devtype,

![is.na](http://is.na)(DevType)

)

survey\_results\_parsed %>%

mutate(Gender = fct\_infreq(Gender)) %>%

ggplot(aes(Gender, ConvertedComp, color = Gender)) +

geom\_boxplot(outlier.colour = NA) +

geom\_jitter(aes(alpha = Gender), width = 0.15) +

facet\_wrap(~DevType) +

scale\_y\_log10(labels = dollar\_format()) +

scale\_alpha\_discrete(range = c(0.04, 0.4)) +

coord\_flip() +

theme(legend.position = "none") +

labs(

x = NULL, y = NULL,

title = "Salary and Gender in the 2019 Stack Overflow Developer Survey",

subtitle = "Annual salaries for US developers"

)

In almost all of these developer role categories, we see that women, and sometime non-binary folks, earn less than men.

survey\_results\_parsed %>%

mutate(Gender = fct\_infreq(Gender)) %>%

group\_by(Gender, DevType, OpenSourcer) %>%

summarise(YearsCodePro = median(YearsCodePro, na.rm = TRUE)) %>%

ungroup() %>%

ggplot(aes(Gender, YearsCodePro, fill = OpenSourcer)) +

geom\_col(position = position\_dodge(preserve = "single")) +

facet\_wrap(~DevType) +

labs(

x = NULL,

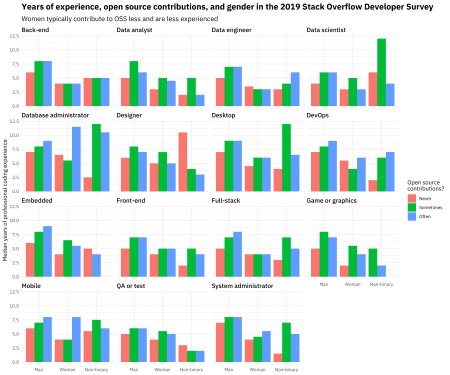
y = "Median years of professional coding experience",

fill = "Open source\ncontributions?",

title = "Years of experience, open source contributions, and gender in the 2019 Stack Overflow Developer Survey",

subtitle = "Women typically contribute to OSS less and are less experienced"

)



Having more years of coding experience tends to be more common for men, for people who contribute more to open source software, and for certain kinds of coders like DBAs and embedded developers.

survey\_results\_parsed %>%

filter(YearsCodePro < 25) %>%

mutate(Gender = fct\_infreq(Gender)) %>%

group\_by(Gender,

YearsCodePro = 5 \* YearsCodePro %/% 5

) %>%

summarise(

ConvertedComp = median(ConvertedComp, na.rm = TRUE),

Total = n()

) %>%

ungroup() %>%

filter(Total > 100) %>%

ggplot(aes(YearsCodePro, ConvertedComp, color = Gender)) +

geom\_line(size = 1.5, alpha = 0.8) +

scale\_y\_continuous(

labels = scales::dollar\_format(),

limits = c(0, NA)

) +

labs(

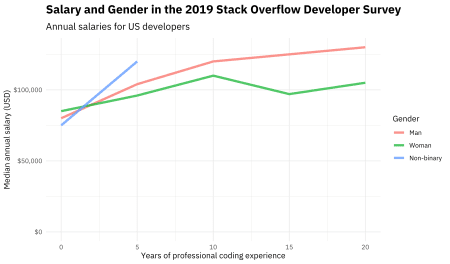
x = "Years of professional coding experience",

y = "Median annual salary (USD)",

title = "Salary and Gender in the 2019 Stack Overflow Developer Survey",

subtitle = "Annual salaries for US developers"

)



Here we see the overall sample-wide salary trends with gender and experience. Among the least experienced developers with fewer than five years of professional experience, women report higher compensation than their counterparts, but the story is different for all other experience cohorts. When we build models, we would like to take this kind of relationship into account.

Also, unfortunately, notice that there are very few non-binary respondents in the higher experience bins. There are so few that I could not fit reliable models in the next sections when including this gender category. I am pretty disappointed about this, because it is so common to see papers and blog posts say, “We know gender isn’t a binary but we’re treating it that way anyway!!!” and that is quite concerning. However, given the realities of the dataset I have, the modeling sections below focus on only the two most represented gender categories in this dataset.

**Building some models**

I can only fit reliable models for salaries up to about $300k USD, for these US salaries. That’s still a pretty impressive range and I’m happy with that, but do note that some respondents reported salaries above that range, as you can see from the exploratory plots above.

modeling\_df <- survey\_results\_parsed %>%

filter(

ConvertedComp < 3e5,

YearsCodePro < 30

) %>%

filter(Gender %in% c("Man", "Woman")) %>%

select(-Respondent) %>%

mutate(ConvertedComp = log(ConvertedComp))

simple1 <- lm(ConvertedComp ~ 0 + DevType + ., data = modeling\_df)

summary(simple1)

##

## Call:

## lm(formula = ConvertedComp ~ 0 + DevType + ., data = modeling\_df)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1.53364 -0.22250 -0.01181 0.21356 1.27291

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## DevTypeBack-end 11.1854712 0.0084200 1328.433 < 2e-16 \*\*\*

## DevTypeData analyst 11.0152280 0.0150496 731.926 < 2e-16 \*\*\*

## DevTypeData engineer 11.2169912 0.0150064 747.481 < 2e-16 \*\*\*

## DevTypeData scientist 11.2066056 0.0173951 644.238 < 2e-16 \*\*\*

## DevTypeDatabase administrator 11.0419249 0.0130809 844.128 < 2e-16 \*\*\*

## DevTypeDesigner 10.9849691 0.0140890 779.682 < 2e-16 \*\*\*

## DevTypeDesktop 11.0972921 0.0105021 1056.673 < 2e-16 \*\*\*

## DevTypeDevOps 11.2359004 0.0110434 1017.429 < 2e-16 \*\*\*

## DevTypeEmbedded 11.1358662 0.0151228 736.363 < 2e-16 \*\*\*

## DevTypeFront-end 11.1003432 0.0090573 1225.568 < 2e-16 \*\*\*

## DevTypeFull-stack 11.1392742 0.0080045 1391.626 < 2e-16 \*\*\*

## DevTypeGame or graphics 11.1043672 0.0219672 505.497 < 2e-16 \*\*\*

## DevTypeMobile 11.1619153 0.0121835 916.150 < 2e-16 \*\*\*

## DevTypeQA or test 11.0400873 0.0139732 790.088 < 2e-16 \*\*\*

## DevTypeSystem administrator 11.0738737 0.0138475 799.701 < 2e-16 \*\*\*

## EdLevelBachelor's degree 0.1321568 0.0057841 22.848 < 2e-16 \*\*\*

## EdLevelGraduate degree 0.2405239 0.0076622 31.391 < 2e-16 \*\*\*

## OpenSourcerSometimes 0.0563180 0.0056426 9.981 < 2e-16 \*\*\*

## OpenSourcerOften 0.1044866 0.0055507 18.824 < 2e-16 \*\*\*

## YearsCodePro 0.0240749 0.0003811 63.175 < 2e-16 \*\*\*

## GenderWoman -0.0261230 0.0081339 -3.212 0.00132 \*\*

## DependentsYes 0.0001854 0.0050973 0.036 0.97098

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3437 on 22828 degrees of freedom

## (303 observations deleted due to missingness)

## Multiple R-squared: 0.9991, Adjusted R-squared: 0.9991

## F-statistic: 1.164e+06 on 22 and 22828 DF, p-value: < 2.2e-16

I fit these models to predict the log of salary (in USD), since it is distributed in a pretty log normal way. There is an intercept for each developer role, and let’s walk through the rest of the predictors:

* A graduate degree is worth almost twice as much as a bachelor’s degree, in terms of the increase in salary.
* Contributing to open source more is associated with higher salary, but the effect size is smaller than, say, a bachelor’s degree.
* Each additional year of experience increases salary (in this linear model, because that’s what linear models do).
* Women earn less than men.
* Having dependents does not affect salary.

The average difference between women’s and men’s salaries, controlling for all these characteristics like developer role, education, experience, and open source contribution, is about the same as a year’s worth of experience.

Let’s try adding an interaction term, to reflect the relationship we observed between salary, gender, and experience.

simple2 <- lm(ConvertedComp ~ 0 + DevType +

Gender \* YearsCodePro + ., data = modeling\_df)

summary(simple2)

##

## Call:

## lm(formula = ConvertedComp ~ 0 + DevType + Gender \* YearsCodePro +

## ., data = modeling\_df)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1.54038 -0.22337 -0.01248 0.21329 1.27664

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## DevTypeBack-end 11.1784964 0.0084544 1322.213 < 2e-16 \*\*\*

## DevTypeData analyst 11.0086407 0.0150521 731.368 < 2e-16 \*\*\*

## DevTypeData engineer 11.2100528 0.0150115 746.764 < 2e-16 \*\*\*

## DevTypeData scientist 11.1987282 0.0173999 643.610 < 2e-16 \*\*\*

## DevTypeDatabase administrator 11.0359560 0.0130847 843.422 < 2e-16 \*\*\*

## DevTypeDesigner 10.9800669 0.0140835 779.642 < 2e-16 \*\*\*

## DevTypeDesktop 11.0905450 0.0105222 1054.011 < 2e-16 \*\*\*

## DevTypeDevOps 11.2288234 0.0110644 1014.861 < 2e-16 \*\*\*

## DevTypeEmbedded 11.1286171 0.0151299 735.540 < 2e-16 \*\*\*

## DevTypeFront-end 11.0933518 0.0090878 1220.690 < 2e-16 \*\*\*

## DevTypeFull-stack 11.1322142 0.0080429 1384.097 < 2e-16 \*\*\*

## DevTypeGame or graphics 11.0972449 0.0219558 505.436 < 2e-16 \*\*\*

## DevTypeMobile 11.1552506 0.0121959 914.672 < 2e-16 \*\*\*

## DevTypeQA or test 11.0340666 0.0139749 789.564 < 2e-16 \*\*\*

## DevTypeSystem administrator 11.0675899 0.0138514 799.025 < 2e-16 \*\*\*

## GenderWoman 0.0417359 0.0117905 3.540 0.000401 \*\*\*

## YearsCodePro 0.0248836 0.0003940 63.163 < 2e-16 \*\*\*

## EdLevelBachelor's degree 0.1329935 0.0057772 23.020 < 2e-16 \*\*\*

## EdLevelGraduate degree 0.2413286 0.0076524 31.536 < 2e-16 \*\*\*

## OpenSourcerSometimes 0.0557397 0.0056354 9.891 < 2e-16 \*\*\*

## OpenSourcerOften 0.1037107 0.0055441 18.707 < 2e-16 \*\*\*

## DependentsYes -0.0006271 0.0050914 -0.123 0.901969

## GenderWoman:YearsCodePro -0.0105867 0.0013333 -7.940 2.11e-15 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3432 on 22827 degrees of freedom

## (303 observations deleted due to missingness)

## Multiple R-squared: 0.9991, Adjusted R-squared: 0.9991

## F-statistic: 1.117e+06 on 23 and 22827 DF, p-value: < 2.2e-16

If we try modeling an interaction term between gender and years of experience, we *no longer* see a decrease in salary for women overall (in fact, at zero years of experience women earn more) but instead, being a woman reduces the amount that each year of experience is worth. From this model, we would estimate that each additional year of experience is worth about 50% to 60% for a woman what it is for a man. I tried out similar interaction terms for gender plus education, etc, and they were clearly important like the gender/experience interaction.

Is the gender/experience interaction really just measuring the impact of having children? Let’s check out what happens with a gender/dependents interaction.

simple3 <- lm(ConvertedComp ~ 0 + DevType +

Gender \* YearsCodePro +

Gender \* Dependents + ., data = modeling\_df)

summary(simple3)

##

## Call:

## lm(formula = ConvertedComp ~ 0 + DevType + Gender \* YearsCodePro +

## Gender \* Dependents + ., data = modeling\_df)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1.53741 -0.22245 -0.01296 0.21264 1.27793

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## DevTypeBack-end 11.177930 0.008455 1322.059 < 2e-16 \*\*\*

## DevTypeData analyst 11.007859 0.015052 731.340 < 2e-16 \*\*\*

## DevTypeData engineer 11.209784 0.015009 746.867 < 2e-16 \*\*\*

## DevTypeData scientist 11.198077 0.017398 643.639 < 2e-16 \*\*\*

## DevTypeDatabase administrator 11.035512 0.013083 843.486 < 2e-16 \*\*\*

## DevTypeDesigner 10.979576 0.014082 779.695 < 2e-16 \*\*\*

## DevTypeDesktop 11.090069 0.010521 1054.036 < 2e-16 \*\*\*

## DevTypeDevOps 11.228074 0.011065 1014.720 < 2e-16 \*\*\*

## DevTypeEmbedded 11.127962 0.015129 735.553 < 2e-16 \*\*\*

## DevTypeFront-end 11.092559 0.009090 1220.314 < 2e-16 \*\*\*

## DevTypeFull-stack 11.131713 0.008043 1383.989 < 2e-16 \*\*\*

## DevTypeGame or graphics 11.096658 0.021953 505.480 < 2e-16 \*\*\*

## DevTypeMobile 11.154689 0.012195 914.684 < 2e-16 \*\*\*

## DevTypeQA or test 11.033487 0.013974 789.590 < 2e-16 \*\*\*

## DevTypeSystem administrator 11.067003 0.013850 799.047 < 2e-16 \*\*\*

## GenderWoman 0.047151 0.011923 3.954 7.69e-05 \*\*\*

## YearsCodePro 0.024759 0.000396 62.519 < 2e-16 \*\*\*

## DependentsYes 0.003583 0.005277 0.679 0.49721

## EdLevelBachelor's degree 0.132874 0.005776 23.003 < 2e-16 \*\*\*

## EdLevelGraduate degree 0.241435 0.007651 31.555 < 2e-16 \*\*\*

## OpenSourcerSometimes 0.055685 0.005634 9.883 < 2e-16 \*\*\*

## OpenSourcerOften 0.103725 0.005543 18.712 < 2e-16 \*\*\*

## GenderWoman:YearsCodePro -0.009140 0.001416 -6.454 1.11e-10 \*\*\*

## GenderWoman:DependentsYes -0.059987 0.019814 -3.027 0.00247 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3432 on 22826 degrees of freedom

## (303 observations deleted due to missingness)

## Multiple R-squared: 0.9991, Adjusted R-squared: 0.9991

## F-statistic: 1.071e+06 on 24 and 22826 DF, p-value: < 2.2e-16

Having dependents does not affect salary in general (i.e. for men) but it **does** reduce salary for women; the decrease is about the same size as two years’ worth of salary increase. Accounting for which respondents do and do not have dependents does not make the experience/gender interaction go away, though; this is evidence that women earn less for each year of experience, even if we account for the wage gap due to children/dependents.

library(broom)

glance(simple1) %>%

mutate(interaction = "none") %>%

bind\_rows(glance(simple2) %>%

mutate(interaction = "gender + experience")) %>%

bind\_rows(glance(simple3) %>%

mutate(interaction = "gender + experience, gender + dependents")) %>%

select(interaction, sigma, logLik, AIC, BIC, df.residual) %>%

kable()

| **interaction** | **sigma** | **logLik** | **AIC** | **BIC** | **df.residual** |
| --- | --- | --- | --- | --- | --- |
| none | 0.3436894 | -8007.553 | 16061.11 | 16245.95 | 22828 |
| gender + experience | 0.3432233 | -7976.040 | 16000.08 | 16192.96 | 22827 |
| gender + experience, gender + dependents | 0.3431619 | -7971.453 | 15992.91 | 16193.82 | 22826 |

With the diverse set of developer types we have in this dataset, this modeling question is probably a better fit for something like a linear mixed effects model, rather than the basic lm() we started with. Some of my favorite resources for getting started with linear mixed effects models are [Page Piccinini’s tutorials](https://pagepiccinini.com/r-course/lesson-6-part-1-linear-mixed-effects-models/) and [this cheat sheet](https://stats.stackexchange.com/questions/13166/rs-lmer-cheat-sheet) on Cross Validated. In this case, we want developer type to be a random effect, while education, open source contribution, years of experience, dependents, and gender will be fixed effects (explanatory variables).

library(lme4)

lmer1 <- lmer(

ConvertedComp ~ (1 | DevType) + EdLevel +

OpenSourcer + YearsCodePro + Dependents + Gender,

data = modeling\_df

)

summary(lmer1)

## Linear mixed model fit by REML ['lmerMod']

## Formula:

## ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer + YearsCodePro +

## Dependents + Gender

## Data: modeling\_df

##

## REML criterion at convergence: 16159.3

##

## Scaled residuals:

## Min 1Q Median 3Q Max

## -4.4578 -0.6480 -0.0351 0.6207 3.7035

##

## Random effects:

## Groups Name Variance Std.Dev.

## DevType (Intercept) 0.005685 0.0754

## Residual 0.118122 0.3437

## Number of obs: 22850, groups: DevType, 15

##

## Fixed effects:

## Estimate Std. Error t value

## (Intercept) 11.1159770 0.0207347 536.105

## EdLevelBachelor's degree 0.1322817 0.0057837 22.872

## EdLevelGraduate degree 0.2409462 0.0076588 31.460

## OpenSourcerSometimes 0.0564272 0.0056424 10.001

## OpenSourcerOften 0.1046647 0.0055502 18.858

## YearsCodePro 0.0240722 0.0003811 63.172

## DependentsYes 0.0001177 0.0050972 0.023

## GenderWoman -0.0262043 0.0081335 -3.222

##

## Correlation of Fixed Effects:

## (Intr) EdLB'd EdLvGd OpnSrS OpnSrO YrsCdP DpndnY

## EdLvlBchl'd -0.229

## EdLvlGrdtdg -0.163 0.563

## OpnSrcrSmtm -0.122 0.031 0.027

## OpnSrcrOftn -0.139 0.085 0.057 0.478

## YearsCodePr -0.114 0.046 -0.016 -0.069 -0.072

## DependntsYs -0.053 0.068 -0.007 0.001 0.004 -0.376

## GenderWoman -0.055 -0.016 -0.050 0.043 0.098 0.056 0.069

Similar to before we see that more education and more open source contributions are associated with higher salary, and that each additional year of experience increases salary while women earn less. Having dependents does not affect salary.

You may be wondering how robust this result is to the threshold on years of experience for respondents to be included in the model. What happens if we model only people who have 15 years of experience or less? Or 5 years of experience or less? There are some modest changes to the fixed effects, but the main fixed effect that changes in these cases is gender; there are strong signals that years of experience and gender interact when effecting salary. We can build such an interaction term into a mixed effects model in a similar way that we did before.

lmer2 <- lmer(

ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer +

YearsCodePro \* Gender + Dependents \* Gender,

data = modeling\_df

)

summary(lmer2)

## Linear mixed model fit by REML ['lmerMod']

## Formula:

## ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer + YearsCodePro \*

## Gender + Dependents \* Gender

## Data: modeling\_df

##

## REML criterion at convergence: 16104.4

##

## Scaled residuals:

## Min 1Q Median 3Q Max

## -4.4756 -0.6489 -0.0380 0.6209 3.7238

##

## Random effects:

## Groups Name Variance Std.Dev.

## DevType (Intercept) 0.005605 0.07487

## Residual 0.117760 0.34316

## Number of obs: 22850, groups: DevType, 15

##

## Fixed effects:

## Estimate Std. Error t value

## (Intercept) 11.108692 0.020621 538.715

## EdLevelBachelor's degree 0.133000 0.005776 23.027

## EdLevelGraduate degree 0.241858 0.007648 31.625

## OpenSourcerSometimes 0.055794 0.005634 9.903

## OpenSourcerOften 0.103902 0.005542 18.746

## YearsCodePro 0.024758 0.000396 62.519

## GenderWoman 0.047173 0.011923 3.956

## DependentsYes 0.003513 0.005277 0.666

## YearsCodePro:GenderWoman -0.009156 0.001416 -6.466

## GenderWoman:DependentsYes -0.059979 0.019814 -3.027

##

## Correlation of Fixed Effects:

## (Intr) EdLB'd EdLvGd OpnSrS OpnSrO YrsCdP GndrWm DpndnY YCP:GW

## EdLvlBchl'd -0.231

## EdLvlGrdtdg -0.164 0.563

## OpnSrcrSmtm -0.122 0.031 0.027

## OpnSrcrOftn -0.139 0.085 0.057 0.478

## YearsCodePr -0.119 0.050 -0.012 -0.069 -0.073

## GenderWoman -0.069 0.001 -0.024 0.019 0.054 0.206

## DependntsYs -0.053 0.063 -0.005 0.001 0.005 -0.381 0.071

## YrsCdPr:GnW 0.035 -0.019 -0.011 0.011 0.017 -0.277 -0.624 0.107

## GndrWmn:DpY 0.009 0.007 -0.005 0.003 -0.001 0.104 -0.150 -0.264 -0.338

Similar to the regular linear model, we no longer see a decrease in salary from gender alone (and being a woman may increase salary at zero experience) but there is evidence for interactions between experience and gender and dependents and gender such that women earn less for each year of experience and when they have dependents. This model estimates that women gain about 60% of the salary increase for each additional year of experience that men gain, and earn about two years’ worth of salary less when they have dependents.

library(broom.mixed)

glance(lmer1) %>%

mutate(interaction = "none") %>%

bind\_rows(glance(lmer2) %>%

mutate(interaction = "gender + experience, gender + dependents")) %>%

select(interaction, sigma, logLik, AIC, BIC, df.residual) %>%

kable()

| **interaction** | **sigma** | **logLik** | **AIC** | **BIC** | **df.residual** |
| --- | --- | --- | --- | --- | --- |
| none | 0.3436895 | -8079.643 | 16179.29 | 16259.65 | 22840 |
| gender + experience, gender + dependents | 0.3431620 | -8052.185 | 16128.37 | 16224.81 | 22838 |

**Bayesian modeling**

For the final models in this post, let’s fit [Bayesian multilevel models](https://cran.r-project.org/web/packages/brms/vignettes/brms_multilevel.pdf) using the **brms** package. This package extends the formula syntax from **lme4** that we just used, so there’s not much to change! First, let’s train a model with no interaction.

library(brms)

options(mc.cores = parallel::detectCores())

fit\_bayes1 <- brm(

ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer +

YearsCodePro + Dependents + Gender,

data = modeling\_df

)

summary(fit\_bayes1)

## Family: gaussian

## Links: mu = identity; sigma = identity

## Formula: ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer + YearsCodePro + Dependents + Gender

## Data: modeling\_df (Number of observations: 22850)

## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;

## total post-warmup samples = 4000

##

## Group-Level Effects:

## ~DevType (Number of levels: 15)

## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS

## sd(Intercept) 0.08 0.02 0.06 0.13 1.00 626 1056

##

## Population-Level Effects:

## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS

## Intercept 11.12 0.02 11.07 11.16 1.01 646

## EdLevelBachelorsdegree 0.13 0.01 0.12 0.14 1.00 3197

## EdLevelGraduatedegree 0.24 0.01 0.23 0.26 1.00 3201

## OpenSourcerSometimes 0.06 0.01 0.05 0.07 1.00 3341

## OpenSourcerOften 0.10 0.01 0.09 0.12 1.00 3616

## YearsCodePro 0.02 0.00 0.02 0.02 1.00 3986

## DependentsYes 0.00 0.01 -0.01 0.01 1.00 4262

## GenderWoman -0.03 0.01 -0.04 -0.01 1.00 3660

## Tail\_ESS

## Intercept 1010

## EdLevelBachelorsdegree 3091

## EdLevelGraduatedegree 2622

## OpenSourcerSometimes 2901

## OpenSourcerOften 2907

## YearsCodePro 2990

## DependentsYes 2414

## GenderWoman 2826

##

## Family Specific Parameters:

## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS

## sigma 0.34 0.00 0.34 0.35 1.00 4099 2875

##

## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample

## is a crude measure of effective sample size, and Rhat is the potential

## scale reduction factor on split chains (at convergence, Rhat = 1).

Second, let’s train a model with an interaction between years of experience and gender and an interaction between dependents and gender.

fit\_bayes2 <- brm(

ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer +

YearsCodePro \* Gender + Dependents \* Gender,

data = modeling\_df

)

summary(fit\_bayes2)

## Family: gaussian

## Links: mu = identity; sigma = identity

## Formula: ConvertedComp ~ (1 | DevType) + EdLevel + OpenSourcer + YearsCodePro \* Gender + Dependents \* Gender

## Data: modeling\_df (Number of observations: 22850)

## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;

## total post-warmup samples = 4000

##

## Group-Level Effects:

## ~DevType (Number of levels: 15)

## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS

## sd(Intercept) 0.08 0.02 0.06 0.13 1.00 761 1505

##

## Population-Level Effects:

## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS

## Intercept 11.11 0.02 11.06 11.15 1.00 668

## EdLevelBachelorsdegree 0.13 0.01 0.12 0.14 1.00 3815

## EdLevelGraduatedegree 0.24 0.01 0.23 0.26 1.00 3670

## OpenSourcerSometimes 0.06 0.01 0.04 0.07 1.00 3678

## OpenSourcerOften 0.10 0.01 0.09 0.12 1.00 3678

## YearsCodePro 0.02 0.00 0.02 0.03 1.00 4133

## GenderWoman 0.05 0.01 0.02 0.07 1.00 3383

## DependentsYes 0.00 0.01 -0.01 0.01 1.00 4335

## YearsCodePro:GenderWoman -0.01 0.00 -0.01 -0.01 1.00 3897

## GenderWoman:DependentsYes -0.06 0.02 -0.10 -0.02 1.00 3560

## Tail\_ESS

## Intercept 1138

## EdLevelBachelorsdegree 3080

## EdLevelGraduatedegree 3141

## OpenSourcerSometimes 3219

## OpenSourcerOften 2971

## YearsCodePro 2957

## GenderWoman 2691

## DependentsYes 2722

## YearsCodePro:GenderWoman 3242

## GenderWoman:DependentsYes 2794

##

## Family Specific Parameters:

## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS

## sigma 0.34 0.00 0.34 0.35 1.00 4884 3304

##

## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample

## is a crude measure of effective sample size, and Rhat is the potential

## scale reduction factor on split chains (at convergence, Rhat = 1).

One of the best things about using Bayesian modeling is that we can build visualizations to see the full distributions of results like these.

library(tidybayes)

fit\_bayes1 %>%

gather\_draws(`b\_.\*`, regex = TRUE) %>%

ungroup() %>%

mutate(Interaction = "No interaction terms") %>%

bind\_rows(fit\_bayes2 %>%

gather\_draws(`b\_.\*`, regex = TRUE) %>%

ungroup() %>%

mutate(Interaction = "Gender interactions")) %>%

mutate(

.variable = str\_remove\_all(.variable, "b\_|EdLevel|Gender|Yes"),

.variable = str\_replace\_all(.variable, "degree", " degree"),

.variable = str\_replace\_all(.variable, "OpenSourcer", "Open Source? "),

.variable = str\_replace\_all(.variable, "YearsCodePro", "Years Coding")

) %>%

filter(.variable != "Intercept") %>%

mutate(

Interaction = fct\_inorder(Interaction),

.variable = fct\_reorder(.variable, .value)

) %>%

ggplot(aes(x = .value, y = .variable, color = Interaction)) +

geom\_vline(xintercept = 0, color = "gray50", size = 1.2, lty = 2, alpha = 0.5) +

geom\_halfeyeh(fill = "gray80") +

stat\_pointintervalh(.width = c(.66, .95)) +

facet\_wrap(~Interaction) +

theme(legend.position = "none") +

labs(

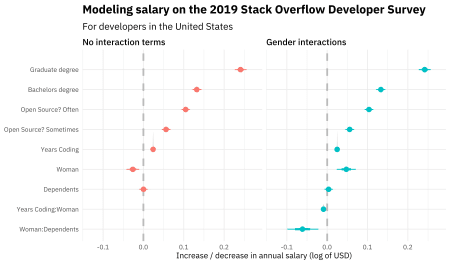
y = NULL,

x = "Increase / decrease in annual salary (log of USD)",

title = "Modeling salary on the 2019 Stack Overflow Developer Survey",

subtitle = "For developers in the United States"

)



I find this style of visualization very informative and helpful. I am a big fan of the [tidybayes](http://mjskay.github.io/tidybayes/) package!

Without interaction terms, being a woman is associated with lower salary; the size of the effect is about the same as one year’s worth of salary increase. Also, having dependents does not affect salary. With an interaction term, women earn more at zero experience, but the interaction term between gender and experience is such that women earn about 2/3 as much for each additional year of experience compared to similar men. Women (but not men) earn less when they have dependents, to the tune of about two years’ worth of salary increases. We can understand this more quantitatively if we create confidence intervals for the model result.

tidyMCMC(

fit\_bayes2,

[conf.int](http://conf.int) = TRUE, conf.level = 0.95,

estimate.method = "median", conf.method = "HPDinterval"

) %>%

mutate(term = str\_remove\_all(term, "b\_")) %>%

filter(!str\_detect(term, "Intercept|sigma")) %>%

arrange(-estimate) %>%

kable(digits = 4)

| **term** | **estimate** | **std.error** | **conf.low** | **conf.high** |
| --- | --- | --- | --- | --- |
| EdLevelGraduatedegree | 0.2418 | 0.0078 | 0.2267 | 0.2566 |
| EdLevelBachelorsdegree | 0.1330 | 0.0058 | 0.1222 | 0.1443 |
| OpenSourcerOften | 0.1038 | 0.0055 | 0.0931 | 0.1147 |
| OpenSourcerSometimes | 0.0557 | 0.0055 | 0.0447 | 0.0664 |
| GenderWoman | 0.0474 | 0.0115 | 0.0240 | 0.0691 |
| YearsCodePro | 0.0247 | 0.0004 | 0.0240 | 0.0255 |
| DependentsYes | 0.0037 | 0.0053 | -0.0065 | 0.0145 |
| YearsCodePro:GenderWoman | -0.0091 | 0.0014 | -0.0117 | -0.0063 |
| GenderWoman:DependentsYes | -0.0602 | 0.0204 | -0.0999 | -0.0207 |

**Discussion**

Based on the salaries reported on the 2019 Stack Overflow Developer Survey, young women (or more precisely, women with very little experience) in the US who code for work appear to be earning a bit higher salaries than similar men, when we control for education, type of work, and open source contributions. The story is not the same if we look at more experienced people who code, where we see evidence that more experienced women earn less for the same work. Depending on the details of how we frame our modeling question, we can estimate the size of the difference by saying that women earn about 2/3 as much for each additional year of experience compared to similar men.

Having dependents does not impact impact salary for men, but it is associated with lower salary for women, by an amount that is about the same as two years’ worth of salary increases. Motherhood is [well known as a big contributor to the gender wage gap](https://www.nytimes.com/2018/02/05/upshot/even-in-family-friendly-scandinavia-mothers-are-paid-less.html) both in the US and around the world. Keep in mind that the professional experience question in this survey was framed as:

How many years have you coded professionally (as a part of your work)?

We can’t guarantee how every person interpreted this question, but hopefully most people did answer this with the years they have worked, not including, say, extensive time off caring for children. I think it is unlikely that a large proportion of this difference can be due to women including years of maternity leave in this response, especially in the US where hardly anyone is lucky enough to have long maternity leave.

So what does this all mean? Remember that this data was collected in 2019; think of it as a snapshot of people with different characteristics (gender, experience, dependent care, type of work) at one point in time. We don’t know what will happen to the folks who are new to our industry today in the next decades. If you an optimist, maybe you see the results presented here as evidence for how much the tech industry has improved over time. Experienced women working today unfortunately have dealt with negative outcomes (including financial outcomes) over their working years, but women at the beginning of their careers now are being treated great. If you are **not** an optimist, you may place these results in the larger picture of research on [retention problems in the tech industry for women](https://www.researchgate.net/publication/268325574_By_RESEARCH_REPORT_The_Athena_Factor_Reversing_the_Brain_Drain_in_Science_Engineering_and_Technology) and other issues around sexism in tech. This dataset alone can’t address all this, but we can place these results within the broader context of research on pay and labor issues in the software industry.

Notice that in this modeling, I *controlled* for respondent characteristics like open source contributions and different types of coding work, but the fact that we observe gender differences in these characteristics is also evidence for the impact of institutional sexism. There are complex reasons why women are, for example, [engaged in DevOps work](https://insights.stackoverflow.com/survey/2019#developer-profile-_-developer-role-and-gender) (currently one of the [highest compensated types of coding work](https://insights.stackoverflow.com/survey/2019#salary)) at low rates, but they are largely related to structural sexism as well. We can explain part of the difference in women’s salaries by women’s lower participation in open source, but that’s… also not good! To learn more about this in particular, you can read detailed results from a [2017 survey on open source](https://opensourcesurvey.org/2017/) organized by GitHub and collaborators.

**Summary**

There are lots of other respondent characteristics not accounted for in this model, including some that are captured in the survey data (how big is the company where someone works? how long have they been in their current job?) and many that are not. In this post, I concentrated on a few factors we know from existing work are important to compensation, but there are lots of further possibilities to explore.

This analysis also focused only on the United States (where I live and work), but the Stack Overflow Developer Survey is an international survey with salary data from around the world. This analysis could be extended to explore these relationships between salary and other characteristics in other countries as well. This may allow a more thorough understanding of the salaries of non-binary developers as well, depending on how consistent the situation is from country to country.

I’m very proud of so much of what I’ve worked on at Stack Overflow, including this extensive annual survey, and I think this analysis demonstrates how important and impactful it is. Let me know if you have any questions or feedback!