Associations of investigation and substantiation proportions to the living wage ratio of California child protective services front line workers

Sofia Guo

December 05, 2023

Contents

I. Context

Background

In California, front-end public child protective services (CPS) workers are required by law to evaluate calls of suspected child maltreatment[1] made by the public to county reporting hotlines. Front-end workers consist of two groups: hotline screeners and emergency response (ER) workers (also known as investigators). The calls (allegations) are evaluated by hotline screeners, based on the state statutory threshold for each type of allegation; the ratio of children with allegations to the estimated child population is the **allegation rate**. Then for the screened-in calls, ER workers are assigned investigations to determine whether the child involved should be removed from their home due to significant evidence of harm or risk to the child; the ratio of children with investigated allegations to the estimated child population is the **investigation rate**[2]. If the ER worker decides that the case demonstrates significant concern or evidence of harm, the allegations are recorded as substantiated; the ratio of children with substantiated investigations to the estimated child population is the **substantiation rate**. For substantiated cases, the ER worker decides what type of intervention is appropriate. These interventions can range from voluntary family services for general neglect cases, to the removal and placement of the child(ren) in foster care for more serious abuse cases; the ratio of children who are placed in foster care to the estimated child population is the **entry rate**.

Purpose

Historically, front-end CPS workers have experienced high rates of burnout and turnover due to chronic understaffing, low pay, and high workloads. Since the pandemic, CPS agencies across the state have seen higher and more persistent vacancies on the front end, which has exacerbated stress on remaining workers and affected the quality of service they provide for each family. Among other reasons, front-end workers highlight the lack of emotional support from management, low pay proportional to their workload and living costs, and feelings of career dissatisfaction as reasons why they leave; these issues are widespread across counties. In response, agencies are looking for ways to increase the retention of their front-end workers. Furthermore, this CPS workforce crisis plays a role in a broader national debate regarding CPS' role in perpetuating racial disparities. Proponents of the UpEnd movement aim to abolish CPS as an institution due to its racially disproportionate involvement in families of color compared to white families. If these workforce issues continue to negatively affect the quality of service and investigations conducted by CPS, agencies may face increased worker errors that cause unnecessary harm to families. By leaving high vacancy and turnover rates unaddressed, agencies could be contributing to higher occurrences of suboptimal outcomes for disadvantaged families – as data show, most families who are referred to CPS are poor

[1] Child maltreatment includes several types of abuse (physical, sexual, emotional, and other) and neglect (general, severe, other) which are defined under California Welfare and Institutions Code Section 300.

[2] In this paper, the investigation and substantiation rates are modified such that instead of total child population as the denominator, the number of children with allegations and investigations are used, respectively. This creates a modified investigation rate that measures the proportion of children with allegations that are investigated, and a modified substantiation rate that measures the proportion of children with investigations that have substantiations. Detailed explanation of this modification is in section II.

families of color. If the goal of CPS is to promote children's safety and well-being in their families, then CPS agencies should strive to improve their workforce conditions such that they provide as much helpful assistance to families as possible and avoid of exacerbating systemic harm. Although this paper only examines investigation and substantiation relationships to the living wage, it is the first step under the broader goal of exploring relationships between county pay, worker outputs, and child outcomes.

Research questions

1. Is the **proportion of children with investigated allegations** (investigation percent) associated with CPS worker annual total wage proportional to the living wage in each county (living wage percent), controlling for county child poverty rate, political conservatism, staff size, and proportion of children with investigations to agency staff size (workload)?
2. Is the **proportion of children with substantiated investigations** (substantiation percent) associated with CPS worker annual total wage proportional to the living wage in each county (living wage percent), controlling for county child poverty rate, political conservatism, staff size, and proportion of children with investigations to agency staff size (workload)?

Motivation

I am interested in examining the effect that investigation and substantiation percents have on the living wage percent because they may point to potential policy solutions. For example, understanding this relationship could point to a policy that increases worker wages to combat higher workloads in a given county, as currently wage increases are not dependent on worker output volume (excluding overtime hours). Each specific question's motivation is detailed below:

1. Differences in investigation percents may be associated with differences in living wage percents because higher investigation percents could be driven by higher baseline maltreatment risks in poorer counties (who cannot pay workers a living wage). Higher investigation percents indicate a higher workload because investigations are very time consuming and require physical visits to the family, which translates to increased stress upon workers who are paid proportionally less than their counterparts in richer counties (which tend to have lower rates of allegations which lowers the overall workload). It may be that ER workers who are paid less and work more are more likely to burn out sooner, leading to higher turnover and overall negative effects on worker capacity to provide quality family services.

2. Differences in substantiation percents may be associated with differences in living wage percents because higher substantiation percents could be driven by higher baseline maltreatment risks in poorer counties (who cannot pay workers a living wage). Higher substantiation percents indicate a higher workload because substantiated cases require worker intervention and can include extensive paperwork such as court reports, which translates to increased stress upon workers who are paid proportionally less than their

counterparts in richer counties (which tend to have lower rates of allegations which lowers the overall workload). It may be that ER workers who are paid less and work more are more likely to burn out sooner, leading to higher turnover and overall negative effects on worker capacity to provide quality family services.

There is sufficient variation across counties in California which allow for meaningful comparison and contrast of different and similar counties. Based on anecdotal evidence, rural counties are more politically conservative than urban counties, while certain large counties by child population tend to have lower child poverty rates and consequently lower allegation rates. The goal of this study is to explore whether such patterns show up in this sample and if other patterns emerge.

Control variables

For both models:

- **Child poverty rate**: Child maltreatment is known to have direct links to child poverty, where poorer children are at higher risk of maltreatment. Poorer counties likely have higher rates of allegations, investigations, and substantiations, and they may pay workers less due to budget constraints. By controlling for child poverty, I avoid confounding the effect of lower wages driven by county impoverishment with the effect of investigation/substantiation percents.
- **County political climate**: More politically conservative counties are known to restrict funding for social services and therefore wages for CPS workers. In addition, they are known to promote a less interventionist approach to families (known as "family preservation" ideology) which may influence workers' propensity to investigate and substantiate cases. By controlling for political climate, I avoid confounding the effect of conservatism on lower wages with the effect of investigation/substantiation percents.
- **County agency size**: Larger agencies are likely to have more funding to hire more workers and may also tend to pay higher wages. By controlling for agency staff size, I avoid confounding the effect of bureaucratic idiosyncrasies on wages with the effect of investigation/substantiation percents.
- **County workload**: Agencies with high ratios of investigations to workers are likely understaffed, resulting in each worker taking on more cases or tasks than average. This relationship may counter any upward pressure on wages related to county size and resource availability. It is possible that both well-funded and underfunded agencies are understaffed due to a widespread front line worker retention crisis throughout CPS in California. Thus, including this control avoids confounding the effect of workload on wages with the effect of investigation/substantiation percents.

II. Method

Sample selection

For all variables of interest, I compiled data for 53[3] out of the 58 total counties in California for the calendar year 2022 because they are the most recent data available and may reflect current conditions of the CPS worker labor force (which are potentially highly variable over the past few years due to COVID). This cross-section of data allows me to first examine my variables of interest without accounting for time, such that I can later expand to a longitudinal analysis once I understand the relationships between my current variables. In this dataset, each row is a unique worker with a county identifier, department name, and position title. Thus, the level-1 unit is a worker nested within a county, while the level-2 unit is a county.

Data sources:

a. Wage data (dollars) The State Controller's Office (SCO) of California requires all counties and cities to report compensation data under sections 53891 and 53892 of the Government code. These data include "pay and benefit information for positions in cities, counties, special districts, and state government... SCO posts the information as it was reported by each public employer, and does not audit for accuracy" (publicpay.ca.gov). The total wage data used in this study is directly downloaded from the "Downloads" tab on the Controller's website. Specifically, I downloaded the 2022 County Data and 2022 City Data csv files. The reason for including the city data file is that some counties are reported as cities, so they are only present in the city file. I filtered the dataset by each county's specific name for its department that contains its CPS agency, in addition to the title of the worker's position that is most likely to be a front-end worker (either hotline or ER worker). These position names were cross-checked by searching each county's website for relevant documentation or job postings and using information from the UC Berkeley MSW program Title IV-E career resources. I also dropped wages that were below the minimum annual wage for each position as reported in the original data ("MinPositionSalary") so that my sample only includes workers who were at least full-time (2,080 work hours per year or 40 hours per week). The units for this measure are in dollars.

b. Living wage data (dollars) Using living wage estimates from livingwage.mit.edu, I navigated to "California" on their main page, then clicked on each corresponding county's link to obtain the hourly living wage in dollars for two adults (one working) who have two children. I chose this value because it is approximately the midpoint in the "distribution" of estimates given by the living wage calculator (where on the low end is one adult working with no children, and the high end is two adults working with three children). I entered this hourly living wage into a spreadsheet by county, then multiplied them by 2,080 hours work hours to obtain an annualized living wage value in dollars. Finally, these data were joined to

[3]The five missing counties are Sierra, Santa Clara, Contra Costa, Mendocino, and Lassen, due to either lack of reporting to the state controller or missing worker wage data for calendar year 2022.

the main wage data such that each row corresponds to the county's estimated annualized living wage (in dollars).

Living wage percent: My final response variable is a computed variable, the living wage percent, which is equal to each worker's annual total wage divided by the annualized living wage estimated for each county, then multiplied by 100 to convert to a percent unit. This number can be interpreted as the percent of the estimated living wage that each worker earned in 2022. The purpose of this transformation is to incorporate information about the cost of living in each county because higher paying counties often have higher living costs (and vice versa). I am interested in learning which workers earn a living wage in each county to more accurately characterize the financial well-being which could indicate lower work stress and better service capacity

c. Investigation/substantiation percents The number of children with allegations, investigations, and substantiations data by county were downloaded from the Child Welfare Indicators Project (CCWIP) secure site[4]. For these data, I pulled the "JAN2022-DEC2022" columns and computed the following proportions for each county:

Investigation percent: This measure equals the number of children with investigations divided by the number of children with allegations times 100. One can interpret this as a percent measure of a county's propensity to investigate allegations (how interventionist it is, given the same maltreatment threshold) or as a measure of a county's child population vulnerability (more investigations means more allegations are identified as concerning). Both interpretations are important to consider because the statutory threshold for maltreatment is broadly defined and subject to discretion. Even with the help of structured decision making tools (SDM) at the front lines, workers are able to override SDM recommendations if they deem it necessary.

Substantiation percent: This measure equals the number of children with substantiated investigations divided by the number of children with investigations times 100. One can interpret this as a percent measure of a county's propensity to substantiate investigations, which, similar to investigation percents, can be interpreted as some combination of county intervention propensity and actual underlying child population vulnerability. This project does not aim to trace this measure's root causes, but it attempts to first explore whether they are even related to wage ratios.

d. Child poverty rates (percent) These data were downloaded from CCWIP under their secure site (see footnote 4). The measure is equal to the estimated number of children in poverty divided by the estimated number of children living in a county multiplied by 100. According to CCWIP, these data are derived from the 5-year estimates from the American

[4]I have access to these data due to my affiliation to CCWIP as a graduate student researcher. This means that the public cannot directly access secure site data or download it for use. More restricted versions of these data are available on the public site (ccwip.berkeley.edu).

Community Survey (ACS) multipliers. The poverty population is the product of the ACS multiplier, while the child population estimates are from department of finance projections.

e. Percent voted for Biden (percent) Using data from Politico, I hand-entered in Excel the percent of a given county which voted for Biden in the 2020 presidential election as a measure for how conservative it is. This value is reported as "BIDEN PCT" in the "County results" portion on the Politico website.

f. Workers per agency (z-score) Using the wage data, I tabulated the number of rows per agency to construct the number of workers in each county (one agency per county) because each row in the wage dataset represents a worker. This measure is equivalent to the number of units per cluster in this project[5]. I then subtract the mean number of workers per county from each observation and divide by the standard deviation of this measure to compute standardized z-score units to make interpretation easier.

g. Investigations to worker ratio (z-score) To capture a measure of workload, I divided the estimated number of children with investigations (from CCWIP) by the number of workers in each county (units equaling children per worker). This should give a general sense of how burdened a given county's workforce is because investigations constitute the subset of allegations that are screened in and require a worker to knock on a family's door, hence more time consuming and heavier work involved (than just screening calls). However, it may not be particularly precise due to uncertainty around how accurately this wage dataset captures worker count. Nevertheless, I compute the z-score for this measure by subtracting the mean ratio of children with investigations to workers for each county and divide by the standard deviation of the measure to make interpretation easier.

III. Models

I estimate two multilevel random intercept models. The first model has $investalleg_j$ as the main independent variable of interest, and the second model estimates the β_2 coefficient for $subinv_j$ instead (all other covariates remaining the same). To be concise, I write out the first model explicitly only. The second model is identical except for the switching of the main independent variable of interest as mentioned above. The model is as follows:

$$lwratio_{ij} = \beta_1 + \beta_2 investalleg_j + \beta_3 childpovrt_j + \beta_4 percbiden_j$$
$$\beta_5 scaledworker_j + \beta_6 scaledcase_j + \zeta_j + \epsilon_{ij}$$

where:

- i is the worker (level 1 unit)
- j is the county that the worker is nested within (level 2 unit)

[5]This may be imprecise because each county may not fully report their wages, but a general summary statistic table shows that most of the large counties seem to be reporting reasonably sized worker counts based on anecdotal evidence.

- $lwratio_{ij}$ is worker i's total annual wages (dollars) divided by the annualized living wage (dollars) for county j multiplied by 100 (living wage percent)
- β_1 is the population mean of the living wage ratio multiplied by 100
- β_2 is the association between the number of children with investigations divided by the number of children with allegations in county j multiplied by 100 (the investigation percent) and the living wage percent[6], controlling for:
  - β_3, the association between the number of children in poverty divided by the total child population in county j multiplied by 100 (child poverty percent) and the living wage percent
  - β_4, the association between the percent of people who voted for Biden in the 2020 election in county j (Biden percent) and the living wage percent
  - β_5, the association between the standardized (z-score) count of workers in county j and the living wage percent
  - β_6, the association between the standardized (z-score) number of children with investigations divided by the number of workers in county j and the living wage percent
- ζ_j is the level-2 residual (deviation of county j's mean living wage percent from the population mean)
- ϵ_{ij} is the level-1 residual (deviation of the living wage percent for worker i from county j mean)

In addition:

- ζ_j is assumed to have a normal distribution with mean zero and variance ψ (between-cluster variance)
- ϵ_{ij} is assumed to have a normal distribution with mean zero and variance θ (within-cluster) variance

My primary parameter of interest is β_2 because I hypothesize that counties with higher investigation and substantiation percents will have lower living wage percents (β_2 will be negative in both models) controlling for child poverty, staff size, investigations to staff ratio, and political conservatism; this hypothesis is based on anecdotal evidence that counties with high investigation and substantiation percents tend to have less resources in general, as most allegations of maltreatment involved child neglect which is highly associated with conditions generated by child poverty. Because CPS worker wages are constrained by the amount of general funding available to a county, poorer counties place a higher labor burden on their workers because they are not able to pay (and thus retain/attract) workers as competitively as richer counties despite having relatively higher proportions of investigations due to higher rates of child poverty. These two parameters should answer my research questions because they estimate how every percentage point increase in the investigation/substantiation percent influences the living wage percent.

[6]For the second model, β_2 is the association between the number of children with substantiated investigations divided by the total number of children with investigations (substantiated plus unsubstantiated) multiplied by 100 (the substantiation percent) and the living wage percent, controlling for the same covariates.

IV. Description

Table 1: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
LWratio	8,157	98.31	21.87	1.05	301.31
ChildPovRt	8,157	14.12	3.63	5.37	25.30
InvAlleg	8,157	71.12	12.15	35.12	86.58
SubInv	8,157	18.61	5.17	7.79	40.45
PercBiden	8,157	63.38	10.38	26.50	85.30
TotalWorkers	8,157	1,794.96	1,634.37	1	3,614
Caseload	8,157	35.80	28.99	11	1,247
scaledworker	8,157	0.00	1.00	−1.10	1.11
scaledcase	8,157	−0.00	1.00	−0.86	41.78

In this sample, there are 53 clusters (counties) and 8,157 level-1 units (workers). The average cluster size is 1,795 workers per county, with a standard deviation of 1,634 workers, a minimum of one worker in a county, and maximum of 3,614 workers in Los Angeles (LA) County. These figures make sense due to the wide variation in county size. For example, LA houses more than two million children, which constitutes nearly 25% of California's child population. Meanwhile, the smallest county Alpine has a child population of 204 children, which amounts to far less than one percent of the state child population, hence the singular worker in a county is also reasonable. This disparity is what drives the large standard deviation in total workers per cluster. The five missing counties (out of 58 total) are Sierra, Santa Clara, Contra Costa, Mendocino, and Lassen, due to either lack of reporting to the state controller or missing worker wage data for calendar year 2022. Santa Clara and Contra Costa are two large counties in the Bay Area region of the state, who had child populations of approximately 500,000 children and 300,000 in 2022, respectively (according to CCWIP). Due to the potentially high number of workers (I estimate around 400-500 total between them based on similar local counties) they employ, these two counties' omission from this estimation sample may influence the estimates somewhat. Sierra, Mendocino, and Lassen counties are much smaller counties with a total child population of roughly 28,000 children combined, so the number of workers they employ are likely much less than 100 in sum. Otherwise, for the estimation sample, there are no missing data for each variable of interest outlined in the models.

For the z-score variables *scaledworker* and *scaledcase* (see section I for details on construction), the respective non-transformed means are 1,795 workers per county with a standard

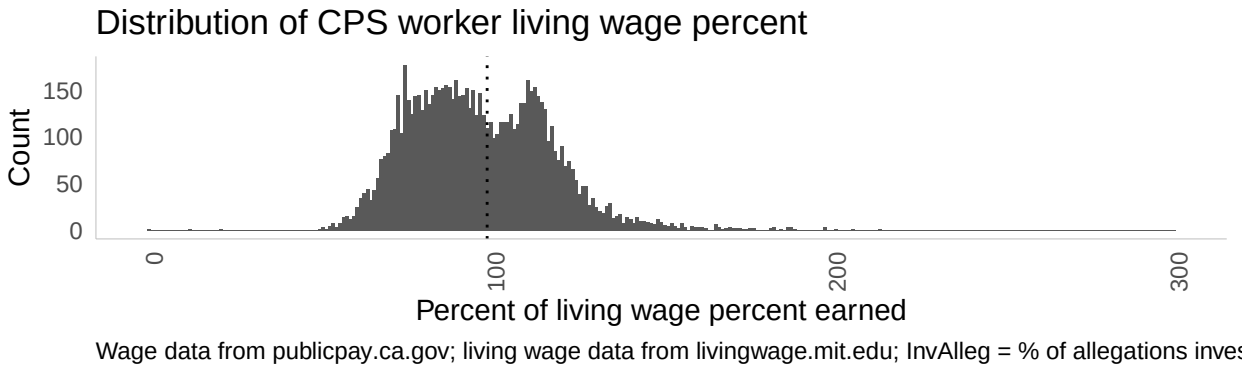
Table 2: Cluster summary statistics: Total workers per county

Mean	SD	Min	Max	Units
1794.96	1634.37	1	3614	8157

Table 3: Overall, between and within-cluster living wage percent statistics

Level	Mean	SD	Min	Max	N
Overall	98.31	21.87	1.05	301.31	8157.00
Between	98.31	13.67	51.57	124.41	53.00
Within	98.31	18.97	0.46	294.13	153.91

deviation of 1,634 workers, while the second variable has a mean of 35 investigations per worker with a standard deviation of 29 investigations. These are represented by the zero value in the summary statistics table, such that a value of -1.1 indicates 1.1 standard deviations below the mean and a value of 42 indicates 42 standard deviations above the mean.



The distribution of the living wage percent response variable seems to be bimodal, with two peaks and a trough centered at 100 percent of the living wage. This indicates that most counties tend to either pay slightly below a living wage (around 80-90 percent), while a smaller group clusters above around 110 percent. Based on a faceted plot of distributions by region and county size, it seems that large southern counties are primarily driving this bimodality (see appendix for figure). Counties outside of that size class and region seem to have relatively normal distributions (excluding small counties).

V. Multilevel modeling

Table 4 presents estimates for the random intercept multilevel models of investigation and substantiation percents, and Table 5 presents the estimated intra-class correlations for each model. Residual diagnostics can be found in section VII (Appendix). Overall, none of the estimated parameters except of the intercepts were statistically significant at the 5 percent level. Thus, we fail to reject the null hypothesis that the covariates of interest have a nonzero association with the living wage percent.

Intercepts

The null model intercept can be interpreted as the estimated mean living wage percent across counties, which was estimated at 95.5 percent and statistically significant at the 0.1 percent level. This is similar to the level-1 overall average of 98.31 percent. The near-100 percent

Table 4: Random intercept linear multilevel model regressions of living wage percent

	Null Model	Model 1	Model 2
Intercept	95.60***	103.67***	89.00***
	(1.86)	(16.07)	(15.54)
Investigation %		−0.20	
		(0.16)	
Substantiation %			0.05
			(0.24)
Child Poverty %		−0.48	−0.65
		(0.36)	(0.34)
Voted Biden %		0.25	0.28
		(0.13)	(0.15)
No. Workers/Agency		3.26	0.68
		(5.31)	(5.05)
No. Investigations/Worker		0.50	0.43
		(0.43)	(0.44)
AIC	71363.04	71356.66	71358.32
BIC	71384.06	71412.72	71414.38
Log Likelihood	−35678.52	−35670.33	−35671.16
Num. obs.	8157	8157	8157
Num. groups: EmployerCounty	53	53	53
Var: EmployerCounty (Intercept)	157.85	112.71	116.66
Var: Residual	362.12	362.06	362.07

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5: Intra-class correlations for living wage percent RIM models

Null	Model.1	Model.2
0.3	0.24	0.24

average indicates that most counties are paying a living wage, which is quite surprising given the high cost of living throughout the state. The first model intercept is interpreted as the estimated mean living wage percent across counties for counties with the zero child poverty rates, zero investigation percents, zero constituents voted for Biden, and 1,795 workers/35 children with investigations per worker (the average value for the z-scores *scaledworker* and *scaledcase*); it is estimated at 103 percent of the living wage, which is statistically significant at the 0.1 percent level. The second model intercept is interpreted as the estimated mean living wage percent across counties for counties with the zero child poverty rates, zero substantiation percents, zero constituents voted for Biden, and 1,795 workers/35 children with investigations per worker (the average value for the z-scores *scaledworker* and *scaledcase*); it is estimated at 89 percent of the living wage, which is statistically significant at the 0.1 percent level.

Parameters

My parameters of interest β_2 for each of models 1 and 2 are not significant at the 5 percent level, thus I cannot reject the null hypothesis that the associations between living wage and investigation/substantiation percents are zero. For model 1, β_2 is estimated at -0.2 with standard error 0.16; this can be interpreted as for every percentage point increase in investigation percent, on average the living wage percent decreases by 0.2 percentage points, controlling for other covariates. The negative direction of this estimate supports my original hypothesis that increased investigation percents may be associated with lower living wage percents due to county impoverishment. However, the magnitude of this estimate is small and indistinguishable from zero. For model 2, β_2 is estimated at 0.05 with standard error 0.24; this can be interpreted as for every percentage point increase in the substantiation percent, on average the living wage percent increases by 0.05 percentage points. The positive direction of this estimate is surprising given my initial hypothesis that higher substantiation percents would be associated with lower living wage percents due to high community and county impoverishment. However, the magnitude of this estimate is even smaller than that for investigation percents and statistically indistinguishable from zero.

ICC

I computed the intra-class correlation for each model using the following formula: $ICC = \frac{\psi}{\psi+\theta}$, where ψ is the between cluster variance not explained by the model and θ is the within cluster variance not explained by the model. For both models 1 and 2, approximately 24% of the variance not explained by the model is between counties, which indicates that nearly 75% of the variance not explained by the models is within counties (at the worker level). This makes sense as the models do not have any level-1 covariates other than the response variable; given that worker wages vary significantly from person to person, it is reasonable to think that most of the variance in my response variable is better captured by level-1 rather than level-2 variables.

VI. Discussion

Overall, the estimated results of these models provide preliminary answers to the research questions posed in section I: that investigation and substantiation percents are not associated with living wage percents. These results also indicate that further study is needed which uses worker-level instead of only county level covariates to explain more of the variation in living wage percents. In addition, the lack of statistical significance (large standard errors) across all the estimated parameters except for intercepts indicates that the models have difficulty obtaining precise estimates for each covariate. This is potentially due in part to the fact that all covariates are level-2 and do not capture the individual unit-level variation within counties that the response variable contains.

Despite these limitations, the estimates reveal some interesting patterns. Other estimated control parameters, although all statistically insignificant at the 5 percent level, point to some interesting directional relationships. β_3, the association between child poverty rate and living wage percent, was negative across both models; this is consistent with my hypothesis that counties which have high poverty rates likely have less funding to spend on worker pay.

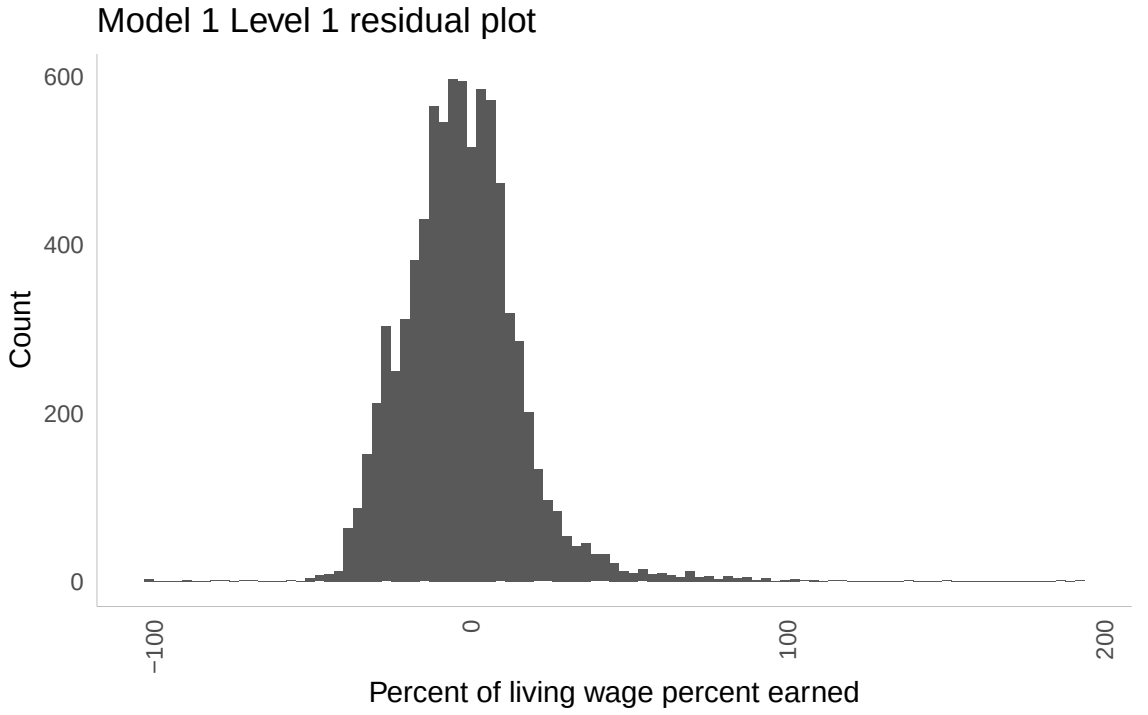
In addition, β_4, the association between percent voting Biden in 2020 and living wage percent, was positive across both models and close in magnitude. This is consistent with my hypothesis that more politically liberal counties spend more on social services and thus pay higher wages to CPS workers.

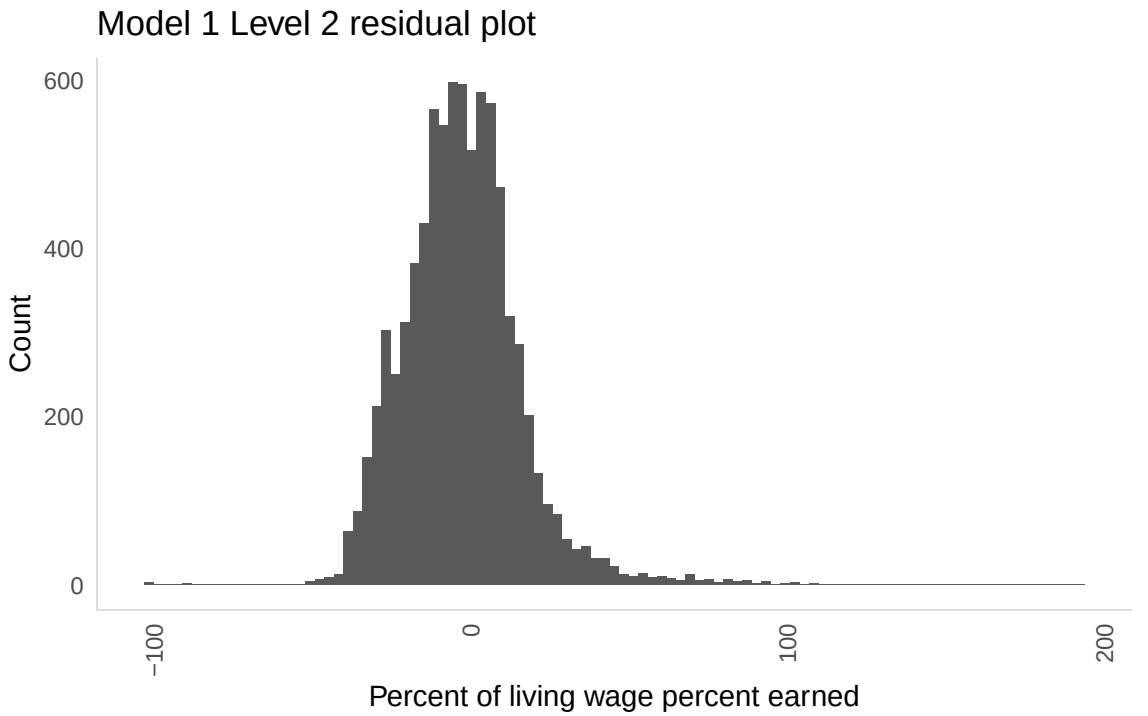
Finally, both β_5 and β_6, the associations between staff size/workload and the living wage percent, respectively, were estimated as positive across both variables and models. The positive association between agency staff size and living wage percent is consistent with my hypothesis that larger agencies have more funding and pay their workers proportionally more. However, I was surprised by the positive association between workload and living wage, albeit small (ranging from 0.5 percentage points to 0.44 percentage points); the positive sign suggests that agencies with higher caseloads pay slightly higher living wage percents. I would like to err on the side of caution in reading into this because of the uncertainty around my construction of the variable (particularly whether row counts are accurate worker counts as reported by the county in the wage dataset).

Further work on this subject should focus on identifying worker-level data that can characterize the financial conditions of the worker, county, and agency. The findings of this project hint that CPS worker wages are likely a proxy for county resource availability in general, and that most counties do pay a living wage. This is an important first step in this area of research because if CPS wages do affect the turnover and retention of front line workers, it would be in the state's best interest to distribute resources to counties with less resources and lower pay. A potential next step in this project will be to look at the association between worker wages and retention rates, controlling for poverty and political climate.

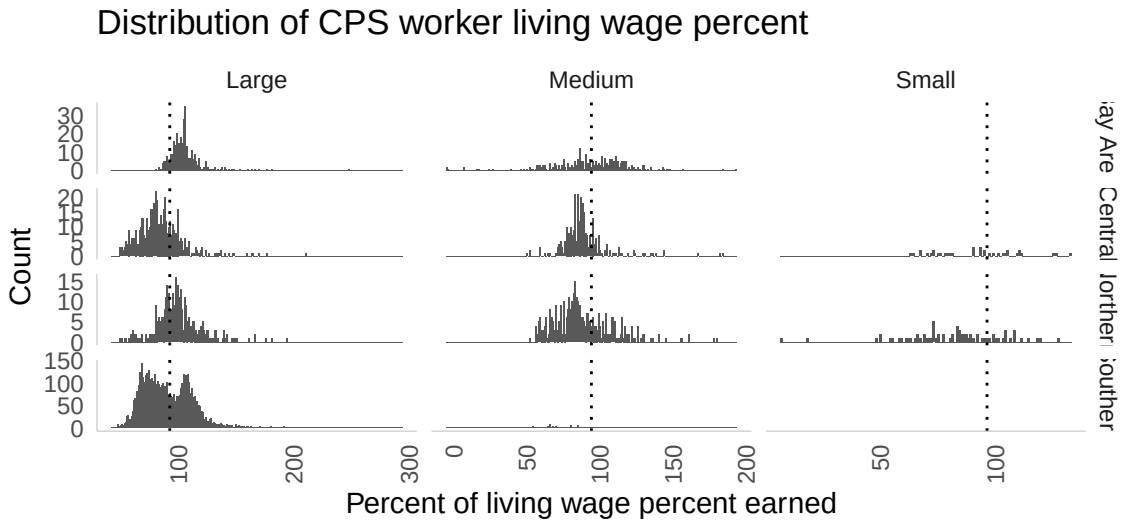
VII. Appendix

Residual diagnostics:

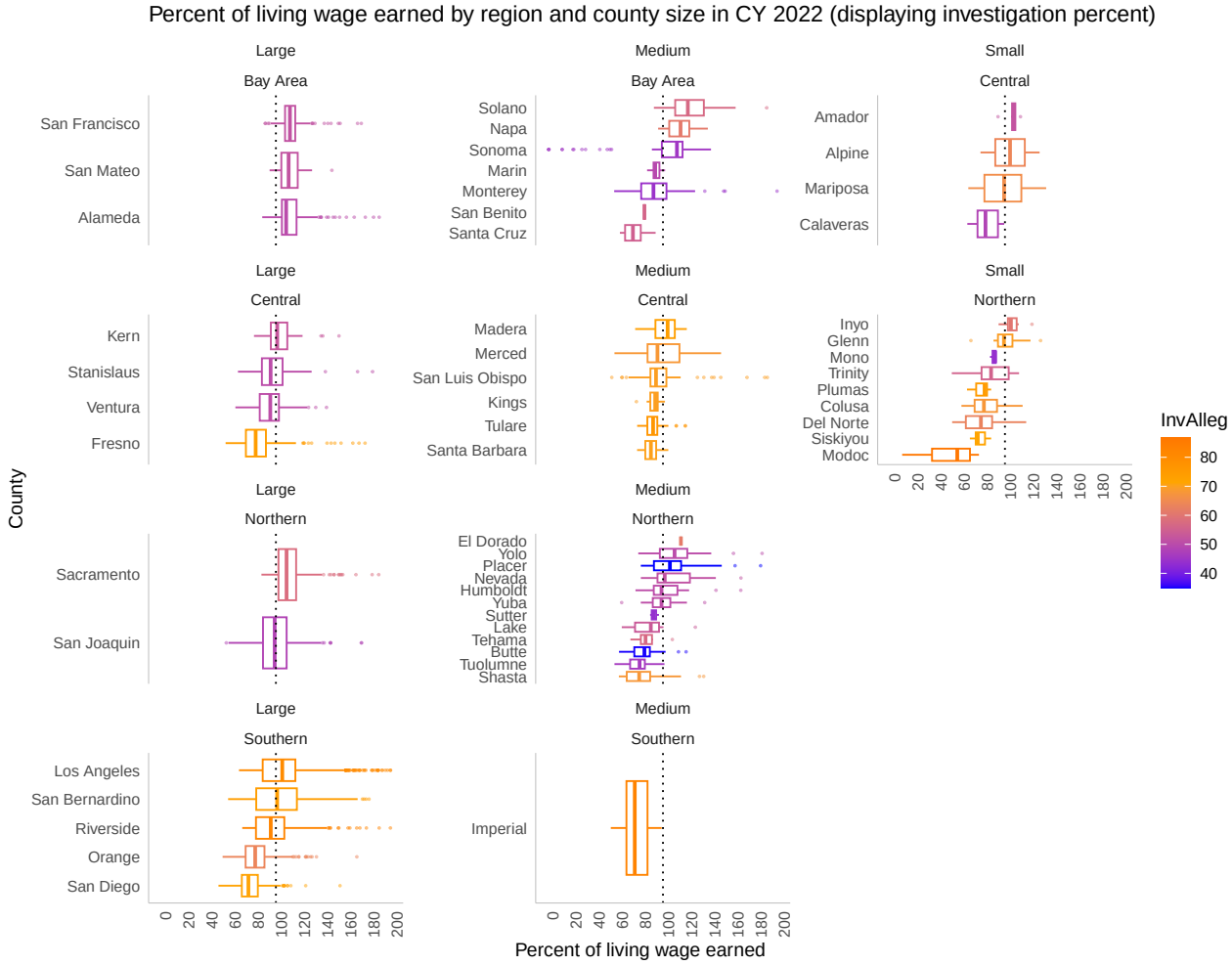




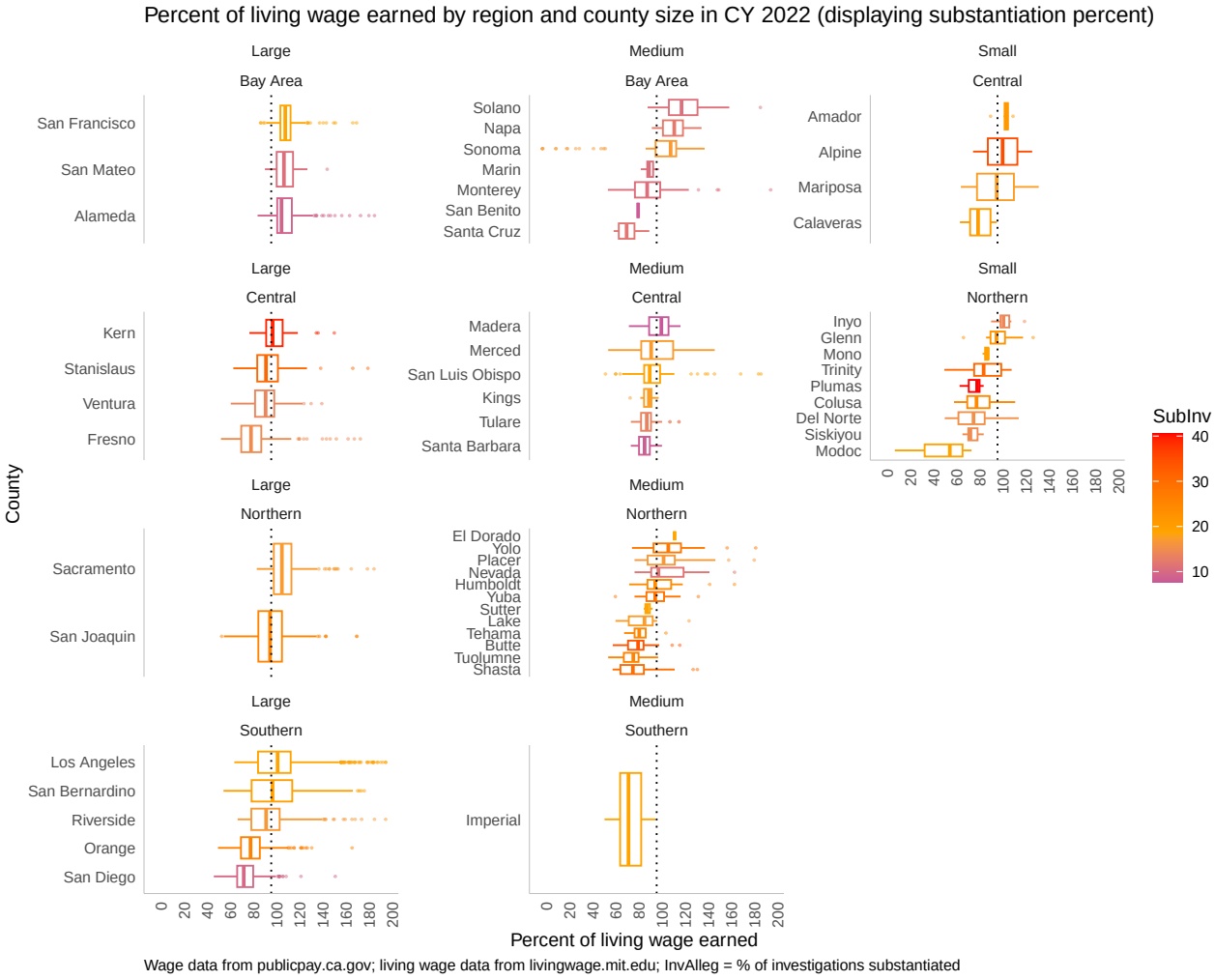
Additional figures:



Wage data from publicpay.ca.gov; living wage data from livingwage.mit.edu; InvAlleg = % of allegati



Wage data from publicpay.ca.gov; living wage data from livingwage.mit.edu; InvAlleg = % of allegations investigated



Stata code:

Because I conducted this analysis in R, I only write the following commands for running the models in Stata (and print the rest of the R code that was used in the subsequent section).

- For model 1: **mixed lwratio InvAlleg ChildPovRt PercBiden scaledworker scaledcase || EmployerCounty: , robust stddeviations**
- For model 2: **mixed lwratio SubInv ChildPovRt PercBiden scaledworker scaledcase || EmployerCounty: , robust stddeviations**

R code:

For portions of the multilevel analysis, I followed Dr. JoonHo Lee's "R Companion for MLMUS4". Based on information in this guide, the **lme4** package I used in this project to run the models estimates robust standard errors. Below is the code I used to generate this report. For data cleaning (creation of the *relevantcounties.xlsx* file), there is a separate and longer R script that I do not include for the purposes of brevity in this report.

```r
#load packages
library(dplyr)
library(openxlsx)
library(reshape2)
library(tidyr)
library(broom)
library(fBasics)
library(stargazer)
library(lubridate)
library(outliers)
library(robust)
library(car)
library(modelr)
library(fRegression)
library(jtools)
library(knitr)
library(kableExtra)
library(readxl)
library(stringr)
library(magrittr)
library(ggplot2)
library(reshape)
library(forcats)
library(scales)

#packages for HLM
library(tidyverse)
library(easystats)
library(haven)
library(labelled)
library(curl)
library(janitor)
library(gtsummary)
library(broom)
library(broom.mixed)
library(ggeffects)
library(multcomp)
library(lmtest)
library(sandwich)
library(contrast)
library(plm)
library(lme4)
library(lmerTest)
library(lmtest)
```

```
library(merTools)
library(pbkrtest)
library(patchwork)
library(clubSandwich)
library(MASS)
library(writexl)
library(texreg)

#read in exported data relevantcounties.xlsx generated by payanalysis.R script (data c
relevantcounties <- as.data.frame(read_excel("/Users/sofia/Library/Mobile Documents/com

#factorize EmployerCounty
relevantcounties$EmployerCounty <- factor(relevantcounties$EmployerCounty)

#create summary statistics table
summarystats <- as.data.frame(relevantcounties[,c("LWratio","ChildPovRt","InvAlleg","Sub

#LaTex output for summary statistics table
stargazer(summarystats,type='latex', header = F, title='Summary statistics', digits = 2)

#calculate cluster
cluster_summary <- relevantcounties %>%
  dplyr::summarise(
    Mean = mean(TotalWorkers),
    SD = sd(TotalWorkers),
    Min = min(TotalWorkers),
    Max = max(TotalWorkers),
    Units = n())

kable(cluster_summary, caption = "Cluster summary statistics: Total workers per county",

# (1) Overall summary
overall_sum <- relevantcounties %>%
  dplyr::summarize(
    Mean = mean(LWratio),
    SD = sd(LWratio),
    Min = min(LWratio),
    Max = max(LWratio),
    N = n()
  )

# (2) Between-group summary
between_sum <- relevantcounties %>%
  group_by(EmployerCounty) %>%
```

```
  dplyr::summarize(
    gr_mean = mean(LWratio),
    n_j = n()
  ) %>%
  ungroup() %>%
  dplyr::summarize(
    Mean = sjstats::weighted_mean(gr_mean, weights = n_j),
    SD = sd(gr_mean, na.rm = TRUE),
    Min = min(gr_mean, na.rm = TRUE),
    Max = max(gr_mean, na.rm = TRUE),
    N = n()
  )

# (3) Within-group summary
df_dev <- relevantcounties %>%
  group_by(EmployerCounty) %>%
  dplyr::summarize(
    gr_mean = mean(LWratio),
    dev = LWratio - gr_mean,
    size = n()
  ) %>% ungroup()

overall_mean <- df_dev %>%
  dplyr::summarize(mean(relevantcounties$LWratio)) %>% pull()

nj_bar <- df_dev %>%
  group_by(EmployerCounty) %>%
  dplyr::summarize(n_j = n()) %>% pull(n_j) %>% mean()

within_sum <- df_dev %>%
  dplyr::summarize(
    Mean = mean(dev, na.rm = TRUE) + overall_mean,
    SD = sd(dev, na.rm = TRUE),
    Min = min(dev, na.rm = TRUE) + overall_mean,
    Max = max(dev, na.rm = TRUE) + overall_mean,
    N = nj_bar)

# Generate and return the final table
tbl_xtsum <- bind_rows(
  overall_sum, between_sum, within_sum
) %>%
  mutate(Level = c("Overall", "Between", "Within")) %>%
  dplyr::select(Level, everything())
```

```r
kable(tbl_xtsum, caption = "Overall, between and within-cluster living wage percent stat

#distribution of living wage
ggplot(relevantcounties, aes(LWratio)) +
  geom_histogram(binwidth=1)+
    geom_vline(xintercept = 100, linetype = "dotted")+
  labs(x = "Percent of living wage percent earned", y="Count",
       title= "Distribution of CPS worker living wage percent",
       caption = "Wage data from publicpay.ca.gov; living wage data from livingwage.mit.
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        axis.text.x = element_text(angle = 90, hjust = 1))

#fit the null model
mod_lmer0 <- lmer(
  formula = LWratio ~ 1 +(1 | EmployerCounty),
  REML = F,
  data = relevantcounties
)

# (1) Fit the random-intercept model using lmer()
mod_lmer1 <- lmer(
  formula = LWratio ~ InvAlleg + ChildPovRt + PercBiden + scaledworker +scaledcase +(1 |
  REML = F,
  data = relevantcounties
)

# (2) Fit the random-intercept model using lmer()
mod_lmer2 <- lmer(
  formula = LWratio ~ SubInv + ChildPovRt + PercBiden + scaledworker +scaledcase +(1 | E
  REML = F,
  data = relevantcounties
)

#view results
texreg(list(mod_lmer0,mod_lmer1,mod_lmer2), caption="Random intercept linear multilevel

# Estimate ICC
M0_ICC <- 157.85/(157.85+362.12)
M1_ICC <- 112.71/(112.71+362.06)
```

```
M2_ICC <- 116.66/(116.66+362.07)

#make ICC dataframe
ICC <- data.frame("Null" = M0_ICC,
                  `Model 1` = M1_ICC,
                  `Model 2` = M2_ICC)

kable(ICC, caption = "Intra-class correlations for living wage percent RIM models", digi

#residuals for lmer1
# Extract level 1 residuals

finaldf <- relevantcounties %>%
  dplyr::mutate(
    M1L1=residuals(mod_lmer1, level=1),
    M2L1=resid(mod_lmer2, level=1),
    M1L2=resid(mod_lmer1, level=2),
    M2L2=resid(mod_lmer2, level=2)
  )

#residual plot level 1
ggplot(finaldf, aes(M1L1)) +
  geom_histogram(binwidth=3)+
  labs(x = "Percent of living wage percent earned", y="Count",
       title= "Model 1 Level 1 residual plot",
       caption = "Wage data from publicpay.ca.gov; living wage data from livingwage.mit.
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        axis.text.x = element_text(angle = 90, hjust = 1))

#residual plot level 2
ggplot(finaldf, aes(M1L2)) +
  geom_histogram(binwidth=3)+
  labs(x = "Percent of living wage percent earned", y="Count",
       title= "Model 1 Level 2 residual plot",
       caption = "Wage data from publicpay.ca.gov; living wage data from livingwage.mit.
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
```

```
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        axis.text.x = element_text(angle = 90, hjust = 1))

#distribution of living wage facet grid
ggplot(relevantcounties, aes(LWratio)) +
  geom_histogram(binwidth=1)+
    geom_vline(xintercept = 100, linetype = "dotted")+
  labs(x = "Percent of living wage percent earned", y="Count",
       title= "Distribution of CPS worker living wage percent",
       caption = "Wage data from publicpay.ca.gov; living wage data from livingwage.mit.
  facet_grid(Region~ChildPopSize, scales = "free") +
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        axis.text.x = element_text(angle = 90, hjust = 1))

#distribution of response variables
# InvAlleg
wages_lwratio_invalleg <- ggplot(relevantcounties, aes(LWratio, fct_reorder(EmployerCou
  geom_boxplot(show.legend = T,
               outlier.size = 0.3,
               outlier.alpha = 0.5)+
  geom_vline(xintercept = 100, linetype = "dotted")+
  #geom_violin(width=1.4)+
  #geom_jitter(color="grey", size=0.2, alpha=0.3) +
  labs(x = "Percent of living wage earned", y="County",
       title= "Percent of living wage earned by region and county size in CY 2022 (displ
       caption = "Wage data from publicpay.ca.gov; living wage data from livingwage.mit.
  facet_wrap(ChildPopSize~Region, scales = "free_y", dir = "v")+
  #facet_grid(ChildPopSize~CAlocation, scales = "free_y", switch = "both") +
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_x_continuous(breaks = seq(0,200,20), limits = range(0,200))+
  scale_color_gradient2(midpoint = mean(relevantcounties$`InvAlleg`), low = "blue", mid
```

```
wages_lwratio_invalleg

#SubInv
wages_lwratio_subinv <- ggplot(relevantcounties, aes(LWratio, fct_reorder(EmployerCounty
  geom_boxplot(show.legend = T,
               outlier.size = 0.3,
               outlier.alpha = 0.5)+
  geom_vline(xintercept = 100, linetype = "dotted")+
  #geom_violin(width=1.4)+
  #geom_jitter(color="grey", size=0.2, alpha=0.3) +
  labs(x = "Percent of living wage earned", y="County",
       title= "Percent of living wage earned by region and county size in CY 2022 (displ
       caption = "Wage data from publicpay.ca.gov; living wage data from livingwage.mit.
  facet_wrap(ChildPopSize~Region, scales = "free_y", dir = "v")+
  #facet_grid(ChildPopSize~CAlocation, scales = "free_y", switch = "both") +
  theme_minimal() +
  theme(plot.caption = element_text(hjust = 0),
        legend.position = "right",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "grey", size = .1),
        axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_x_continuous(breaks = seq(0,200,20), limits = range(0,200))+
  scale_color_gradient2(midpoint = mean(relevantcounties$`SubInv`), low = "blue", mid =

wages_lwratio_subinv
```