From Baseline to Structural Break Model: Inference for Weekly Unemployment Insurance Claims*

Analysis of California Unemployment Data from 1987 to 2023

Nora Adadurova

December 10, 2025

Abstract

This paper studies weekly unemployment insurance activity in California from 1987 to 2023, with the goal of understanding how new unemployment claims respond to labor market conditions over time. Starting from a baseline linear model with same-period explanatory variables, the analysis shows that same-week unemployment measures are insufficient due to strong time dependence and structural instability. Incorporating lagged dynamics, seasonal effects, and structural break indicators, especially for the COVID-19 period, substantially improves model fit and interpretation. The results highlight strong persistence in unemployment insurance claims and demonstrate that major economic shocks play a dominant role in shaping observed patterns. These findings provide insight into the delayed and sustained response of the unemployment insurance system during economic downturns.

Introduction

California's unemployment insurance system serves as a crucial indicator of labor market conditions, offering consistent weekly insights into employment health and stress levels. Each week, the state tracks how many workers file new claims and continue to receive benefits, thereby providing a clear measure of insured unemployment. Historically, this measure has spiked during major economic downturns, including the early 1990s recession (Gardner 1994),

*Project repository available at: https://github.com/AdaNora22/final_project_Math-261A/tree/main.

the Great Recession of 2008 (Federal Reserve History 2013; U.S. Bureau of Labor Statistics 2012), and the COVID-19 crisis in 2020 (Neal and Goodman 2021; Walters 2022).

This analysis is particularly relevant not only from an academic perspective but also for new graduates in California preparing to enter the labor market. By understanding how the labor market has responded to past economic shocks, this study offers insights into key indicators of labor market health and the factors influencing its dynamics. In particular, this paper aims to model and draw inference about the weekly insured unemployment rate while accounting for time dependence, seasonality, and structural breaks.

In the sections that follow, we begin by detailing the dataset and data-cleaning process used to prepare the California Unemployment Insurance Weekly Claims data. While the introduction does not list every variable in detail, it is important to note that the analysis focuses on the insured unemployment rate as the target variable, alongside predictors such as initial and continued claims rates, seasonal indicators, and time trends.

Throughout the modeling process, we use model selection criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to compare model fits. Both AIC and BIC are standard statistical measures that balance model complexity against goodness of fit, helping to identify models that achieve a better trade-off between overfitting and explanatory power. Our findings indicate that incorporating these factors significantly improves the model's performance as measured by these criteria.

In summary, this paper first introduces the dataset and exploratory findings, then moves through a series of increasingly sophisticated models. The inclusion of time dependence, seasonal patterns, and structural breaks not only enhances our understanding of the labor market's response to extreme events but also provides practical insights for new graduates and policymakers alike.

Data

The dataset used in this analysis consists of weekly unemployment insurance (UI) activity for the state of California from January 1987 through the end of 2023. Each row represents a single reporting week and includes the number of new unemployment claims ("initial claims"), the number of individuals continuing to receive benefits ("continued claims"), the total number of covered employees, and the insured unemployment rate calculated by the state. These data originate from the California Employment Development Department (EDD), which publishes weekly UI statistics as part of its administrative reporting system. Because the underlying numbers reflect actual counts rather than survey-based estimates, the dataset is generally free of sampling noise and instead captures genuine shifts in labor market conditions, including the recessions of the early 1990s and 2007-2009, as well as the unprecedented shock during the COVID-19 pandemic.

Data Cleaning

Prior to the analysis, necessary data cleaning and variable type transformation weer performed to proporly accommodate time-series modeling. Two columns with contained time data, Filed.week.ended and Reflecting.Week.Ended, were converted to proper Date objects using lubridate. The dataset contained no missing values, though one duplicated row was identified and removed, and an unintended index column created during import was dropped. The raw file also included two categorical variables: Area.Type and Area.Name. These columns contained only a single value ("State" and "California," respectively) since these variables did not contribute any variation or information, they were both dropped.

Moreover, it is important to note that in the thirty seven year period covered by the used data California's labor force has grown substantially (YCharts 2025). Hence, to avoid misleading results, such as potential overall upward trend due to population growth, this paper uses rates rather than raw counts. Weekly initial and continued claims were each divided by the number of covered employees in that week, producing the initial claims rate and continued claims rate. This data transformation provides a comparable scale and ensures that the regression models capture actual changes in economic conditions rather than overall demographic expansion.

Furthermore, to adjust for seasonal and trends analysis, additional variables were created such as the week of the year, month, quarter, and year. These allow the modeling section to incorporate seasonal patterns and long-run trends.

After data cleaning and variable construction, the final dataset consisted of 1,878 weekly observations spanning thirty-seven years of California's unemployment insurance system. Each observation includes the insured unemployment rate, the primary outcome variable in this analysis, along with key predictors such as weekly initial claims rate, continued claims rate, and total covered employment. Additionally, cleaned data included time-based variables created to capture potential seasonal and calendar-related patterns such as filing week of the year, month, quarter, and calendar year. Together, these variables form a complete record of weekly UI activity and labor market health in California and serve as the foundation for the exploratory analysis and inferential modeling that follow.

Exploratory Data Analysis

Before building regression models, it is important to understand the behavior of the key variables and how they evolve over time. The insured unemployment rate, which serves as the response variable in this project, ranges from approximately 1.50% (2005/11/26 and 2019/10/19) to 27.75% (2020/05/02) across the sample.

Figure 1: Weekly insured unemployment rate (IUR) in California from 1987 through 2023 is shown as the black time-series line. The coral shaded regions represent officially recognized U.S. recession periods (1990–91, 2001, 2007–09, and 2020)

A simple time-series plot (Figure 1) immediately highlights the major episodes of economic disruption. The early 1990s recession, the Great Recession, and especially the COVID-19 shutdown period appear as pronounced spikes, with the COVID surge standing out as an extreme outlier relative to the historical pattern. The week ending May 2, 2020 records the highest insured unemployment rate of 27.75% in the used dataset. This level nearly nine times higher than the average before 2020 which was 3.12% according to the provided data and non-ressesion time average of 3.08%.

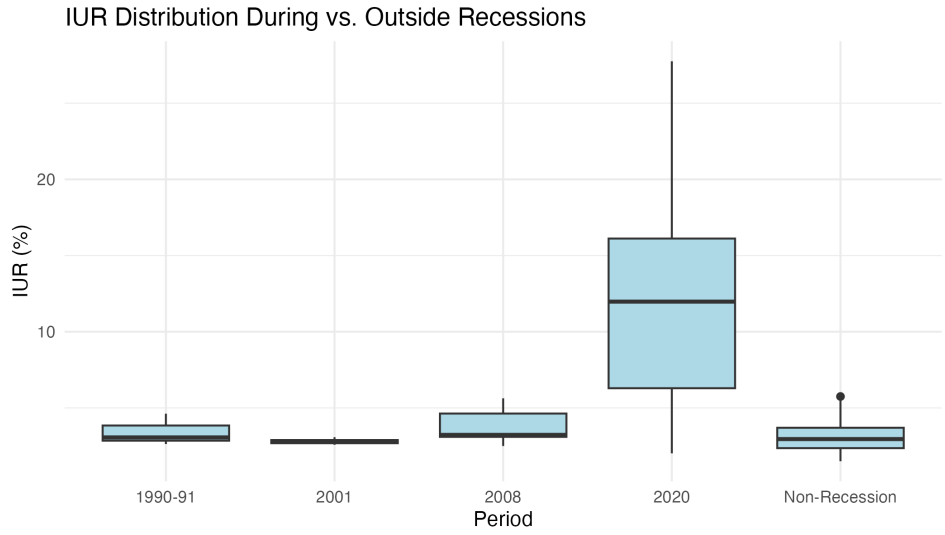

Figure 2: The distribution compareson of the insured unemployment rate across recession periods and non-recession years.

When averaging across periods (Figure 2), the insured unemployment rate increases modestly during the recessions of the early 1990s, 2001, and 2008, but the average rate during the COVID recession exceeds 11%, nearly four times the long-run non-recession average. A comparison of pre- and post-COVID periods further illustrates the structural nature of this change: before March 2020, the maximum insured unemployment rate in the dataset was 5.63%, but after March 2020 the maximum rises to 27.8%.

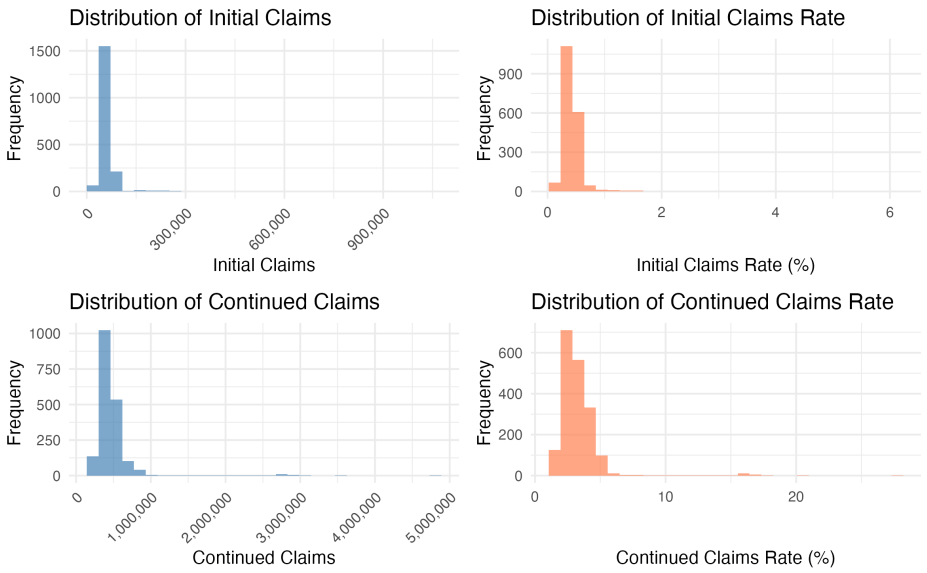

Figure 3: The figure shows the distributions of initial and continued unemployment claims (blue) and their percent rate equivalents (orange).

Histograms of initial claims, continued claims, and their corresponding rates (Figure 3) show strong right skewness. Most weeks fall within moderate ranges, but a small number of economically meaningful extreme events create long upper tails. It is important to note that these are not measurement errors in the data. These outliers reflect real economic shocks. Standard outlier detection using the IQR rule confirms that nearly all outliers arise during recession periods, with COVID-era weeks dominating the extreme tail. Because these values capture the very labor market disruptions this paper aims to study, they were retained in the dataset and later incorporated through structural indicators rather than removed or trimmed.

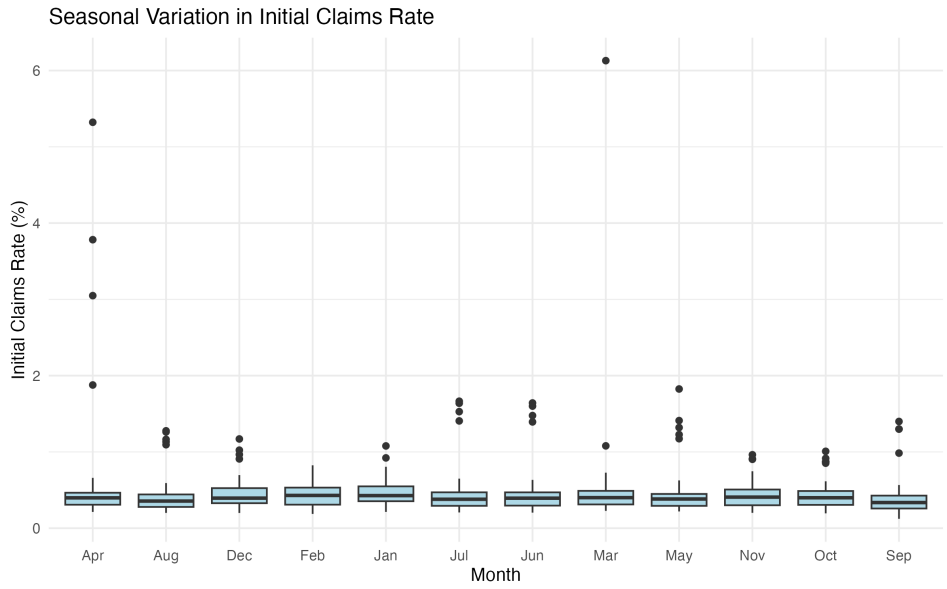

Figure 4: Monthly boxplots of the initial claims rate, illustrating modest but recurring seasonal patterns.

Seasonality is also evident in the exploratory analysis. Monthly boxplots (Figure 4) show that initial claims rates tend to be slightly elevated in January relative to most other months, consistent with recurring post-holiday layoffs and temporary separations. The outliers shown on the plot represent extreme COVID-19 spikes, still the typical cyclical patterns remain detectable, supporting the inclusion of seasonal indicators in later regression models.

The unemployment measures also exhibit strong temporal dependence. The autocorrelation function (ACF) for the insured unemployment rate decays slowly, indicating persistent serial correlation, while the partial autocorrelation function (PACF) shows significant short-lag correlations. Cross-correlations between initial claims rate, continued claims rate, and the insured unemployment rate reveal very strong contemporary relationships, with especially high correlation between the insured unemployment rate and continued claims rate (0.9999). These patterns suggest that including lagged terms is essential for accurately modeling the dynamics of the labor market.

Methods

All modeling was conducted in R [R Core Team (2021)} using base `stats::lm()` for ordinary least squares (OLS), `glmnet` for penalized regression (ridge and lasso), `MASS::rlm()` for robust regression with Huber loss, and `car::vif()` for multicollinearity diagnostics. The goal of this analysis is statistical inference, with an emphasis on understanding which factors are

associated with changes in UI activity while accounting for time dependence, seasonality, and major economic shocks.

A simple regression model that relates the weekly initial claims rate to contemporaneous measures of insured unemployment and continued claims used as an baseline model for understanding how different components of UI activity move together and serves as a reference point for more complex models. However, because the data are weekly and span multiple decades, several classical linear regression assumptions are violated in this baseline setting. In particular, the residuals exhibit strong serial correlation, indicating that current unemployment claims depend heavily on past values. In addition, the presence of large economic shocks, most notably during recessions and the COVID-19 period, introduces heteroskedasticity and structural instability. These features motivate the introduction of lagged variables, seasonal indicators, time trends, and structural break indicators in subsequent models.

To compare the models, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used. They are a model selection criteria that measure how well a statistical model balances goodness of fit with model complexity. Both criteria start from the idea that a model should explain the data well without being unnecessarily complicated. Adding more predictors will almost always improve fit, but overly complex models risk overfitting and poor generalization. AIC and BIC formalize this trade-off. Hence, these two esures are appropriate for the model comparesent for this paper.

Baseline Model

As a starting point, a simple linear regression model is used to study the same-week relationship between the initial UI claims rate and overall labor market conditions. The baseline model includes the insured unemployment rate and the continued claims rate as explanatory variables. Together, these variables summarize the level of unemployment and the extent to which workers remain on unemployment insurance. This specification serves as a reference point before introducing more complex time-series features.

The baseline model is written as

$$Y_t = \beta_0 + \beta_1 U_t + \beta_2 C_t + \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

t indexes weeks and Y_t denotes the weekly initial unemployment insurance claims rate, defined as the number of new claims divided by the number of covered employees. The variables U_t and C_t represent the insured unemployment rate and the continued claims rate, respectively, capturing overall labor market conditions. The coefficients β_1 and β_2 measure the association between these unemployment measures and new claims activity, while the error term ε_t captures unexplained variation and is assumed to be independently and identically distributed with mean zero and constant variance.

Diagnostic results show (Figure 5) that this baseline model violates several key assumptions of classical linear regression. The residuals exhibit strong autocorrelation, indicating that unemployment claims are highly persistent over time. In addition, periods of economic stress—especially during recessions and the COVID-19 period—lead to heteroskedasticity, non-normal residuals, and influential observations. There is also severe multicollinearity between the unemployment measures. These issues are expected given the weekly time-series nature of the data and the presence of large macroeconomic shocks.

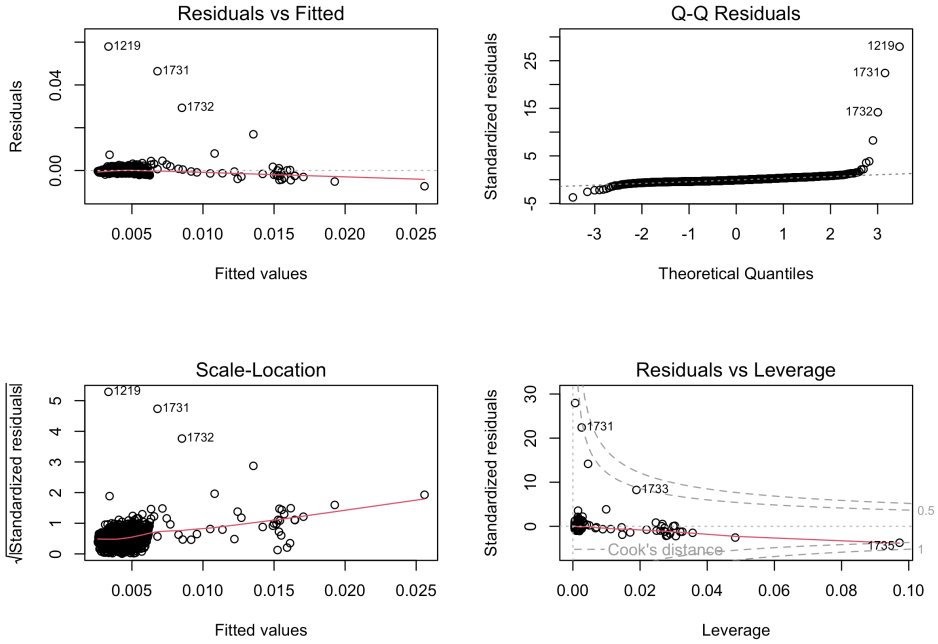

Figure 5: This is the standard OLS diagnostic plots for the baseline regression, showing systematic patterns in the residuals, heavy-tailed deviations from normality, increasing variance with fitted values, and several influential observations.

The standard OLS diagnostic plots indicate that several key linear regression assumptions are violated in the baseline model. Residual diagnostics reveal nonlinearity, strong departures from normality, heteroskedasticity, serial dependence, and influential observations associated with major economic shocks. These violations are expected given the weekly time-series structure of the data and motivate the introduction of lagged variables, seasonal indicators, time trends, and structural break terms in subsequent models.

In terms of fit, the baseline model explains approximately 37% of the variation in the initial claims rate. However, neither explanatory variable is statistically significant, suggesting that same-week labor market measures alone do not adequately explain new claims activity. Al-

though the overall model is statistically significant, this result is driven largely by strong time dependence in the data. The AIC and BIC values (AIC = −17,871.84; BIC = −17,849.69) indicate that while the baseline model provides a reasonable starting point, it is too simple to capture the dynamic behavior of unemployment claims. This motivates the introduction of lagged variables, seasonality, time trends, and structural break indicators in subsequent models.

Advanced Models

To address the strong time dependence observed in the baseline model, an advanced regression specification is introduced that explicitly incorporates lagged information. Unemployment insurance claims evolve gradually over time, and current claims activity is strongly influenced by conditions in prior weeks. This model therefore extends the baseline framework by allowing past values of unemployment indicators to explain current initial claims activity.

Formally, the advanced model is given by

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 U_{t-1} + \beta_4 U_{t-2} + \beta_5 C_{t-1} + \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Here, t indexes weeks. The response variable Y_t denotes the weekly initial unemployment insurance claims rate, defined as the number of new claims filed in week t divided by covered employment. The terms Y_{t-1} and Y_{t-2} capture persistence in new claims activity across recent weeks. The variables U_{t-1} and U_{t-2} represent lagged insured unemployment rates, reflecting broader labor market conditions, while C_{t-1} denotes the lagged continued claims rate, capturing ongoing benefit receipt among unemployed workers. The intercept β_0 represents the baseline level of new claims when all explanatory variables are zero, and the error term ε_t captures unexplained variation. The coefficients β_1 and β_2 measure short-run persistence in initial unemployment claims by capturing how claims activity in the previous one and two weeks affects current claims. The parameters β_3 and β_4 quantify the association between lagged insured unemployment conditions and new claims, reflecting how broader labor market stress translates into future inflows of unemployment insurance applications. The coefficient β_5 captures the contribution of continued benefit receipt to subsequent initial claims activity, linking ongoing unemployment to new separations. Together, these parameters describe how recent labor market conditions propagate forward in time to shape weekly unemployment insurance dynamics.

By incorporating lagged terms, this model directly addresses the autocorrelation that violated key assumptions of the baseline regression. Diagnostic checks indicate substantial improvements in residual independence and overall model fit, as AIC score goes to -18028.17 and BIC score goes to -17989.41, although it is important to note that some deviations from ideal assumptions remain during periods of extreme economic stress. This lagged specification serves as the foundation for subsequent extensions, including the addition of seasonal effects, time trends, and structural break indicators.

Further Improvment of Advance Model

Building on the lagged regression framework, several extensions were explored to further improve model adequacy and assess robustness to alternative specifications. First, seasonal indicators were added to account for recurring calendar effects in unemployment claims, such as elevated layoffs at the beginning of the year. While the inclusion of monthly indicators produced modest improvements in fit, information criteria suggested that seasonality alone does not substantially improve explanatory power once lagged dynamics are accounted for.

Next, a linear time trend was introduced to capture long-run structural changes in the unemployment insurance system, including gradual shifts in labor market participation and policy environments. The addition of a trend term resulted in small improvements in model fit, indicating the presence of slow-moving changes not fully captured by lagged variables alone. Models combining both seasonality and trend terms were also considered, but these more complex specifications were penalized by BIC, suggesting diminishing returns from additional parameters.

To account for major macroeconomic disruptions, indicator variables were added for recession periods and the COVID-19 shock. These structural break terms produced the largest improvements in AIC and BIC among the standard OLS specifications, highlighting the importance of explicitly modeling extraordinary labor market events. The resulting model captures both short-run persistence and abrupt shifts associated with severe economic downturns.

Several alternative estimation strategies were also explored. A log transformation of the response was considered to address skewness and stabilize variance, but this specification performed poorly in terms of information criteria and interpretability. Winsorization was applied to reduce the influence of extreme observations, yielding substantially lower AIC and BIC values; however, this approach intentionally dampens the very shocks of interest and is therefore not appropriate for the primary inferential goals of the analysis. Weighted least squares and robust regression were also estimated to address heteroskedasticity and outliers. While these models improved fit numerically, their information criteria are not directly comparable to standard OLS models due to altered error structures.

A summary of all model comparisons using AIC and BIC is provided in the table below:

Table 1: Model comparison summary using AIC and BIC

Model	AIC	BIC
Baseline	-17871.84	-17849.69
Advanced	-18028.17	-17989.41
Advanced + Seasonality	-18026.19	-17926.53
Advanced + Trend	-18037.33	-17993.04
Advanced + Seasonality + Trend	-18035.89	-17930.69
Advanced + Log	-812.56	-707.36

Table 1: Model comparison summary using AIC and BIC

Model	AIC	BIC
Advanced + Breaks	-18124.89	-18008.61
Winsorized Advanced + Breaks	-21783.83	-21667.55
WLS Advanced + Breaks	-25179.48	-25063.21
Robust Regression	-17967.85	-17851.57

Model Slection

Among the standard OLS specifications, the advanced model with lagged terms and structural break indicators is selected as the preferred model. This specification achieves the lowest AIC and BIC among comparable OLS models while maintaining interpretability and alignment with the inferential objectives of the study. By incorporating lagged unemployment dynamics and explicit indicators for major economic disruptions, the model effectively balances goodness of fit with parsimony and provides a meaningful description of unemployment insurance claim behavior over time.

Although weighted and winsorized models yield substantially lower information criteria, they do so by modifying the error structure or suppressing extreme observations. As a result, these models are not directly comparable to the OLS specifications and are better viewed as robustness checks rather than primary models.

Results

The central research question of this paper is how weekly UI activity responds to labor market conditions once time dependence, seasonality, and major economic shocks are taken into account.

Baseline Results

As an initial reference point, a simple linear regression relates the weekly initial unemployment insurance claims rate to contemporaneous measures of insured unemployment and continued claims. While this baseline model explains approximately 37% of the variation in initial claims, neither explanatory variable is statistically significant. Diagnostic plots reveal strong violations of classical linear regression assumptions, including serial correlation, heteroskedasticity, and non-normal residuals.

Furthermore, figure 6 further illustrates the strong temporal dependence in the baseline model.

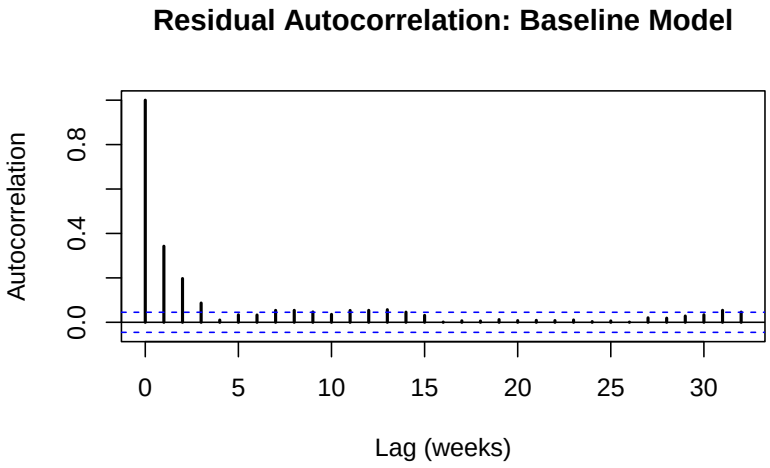

Figure 6: Autocorrelation function of baseline model residuals, showing strong serial dependence.

These results indicate that same-week labor market indicators provide limited information about new claims activity and that unemployment insurance dynamics are strongly time dependent.

Advanced Lagged Model Results

To address the shortcomings of the baseline specification, an advanced regression model incorporating lagged values of unemployment indicators is estimated. Including one and two week lags of the initial claims rate substantially improves model fit, confirming strong persistence in unemployment insurance activity. Lagged insured unemployment and continued claims rates are also associated with current initial claims, indicating that broader labor market stress translates into future inflows of new UI applications.

Compared to the baseline model, the lagged specification achieves a substantially lower AIC ($-18{,}028.17$) and BIC ($-17{,}989.41$), reflecting improved explanatory power after accounting for temporal dependence. Residual diagnostics show a clear reduction in serial correlation relative to the baseline model, although some departures from ideal assumptions remain during periods of extreme economic stress.

Figure Figure 7 compares residual autocorrelation after introducing lagged terms. The advanced model substantially reduces serial dependence, indicating improved adherence to the independence assumption.

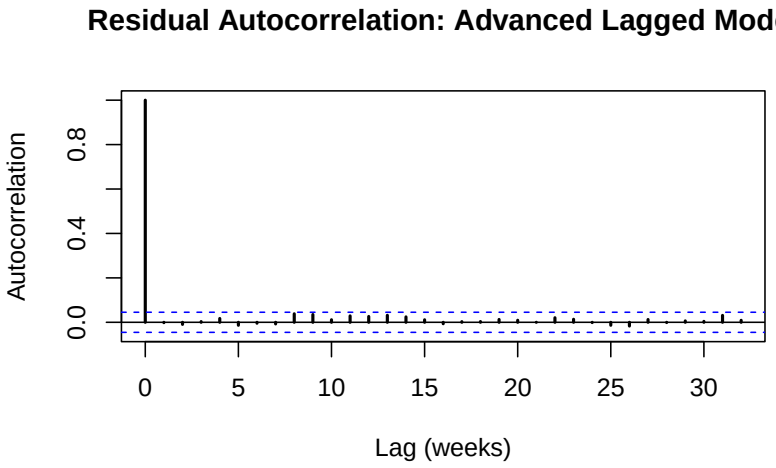

Figure 7: Autocorrelation function of residuals from the advanced lagged model, showing reduced serial dependence relative to the baseline specification.

Extended Model Comparisons

Several extensions of the advanced model were evaluated to improve model adequacy. Seasonal indicators capture modest recurring patterns in UI claims, a linear time trend accounts for long-run changes in the labor market, and structural break indicators for the Great Recession and the COVID-19 period capture sharp regime shifts that are not well explained by smooth dynamics alone.

According to Table 1, among models estimated on the original scale of the data, the specification that includes lagged terms, seasonality, a time trend, and structural break indicators performs best. While the weighted least squares and winsorized models produce much lower information-criterion values, these improvements mainly reflect changes to the error structure and are not directly comparable to standard OLS models. Robust regression yields similar qualitative conclusions but does not materially improve fit relative to the selected specification.

Overall, the results show that unemployment insurance claims are highly persistent and respond to lagged labor market conditions, with major economic shocks, especially the COVID-19 period, playing a dominant role in shaping the observed dynamics.

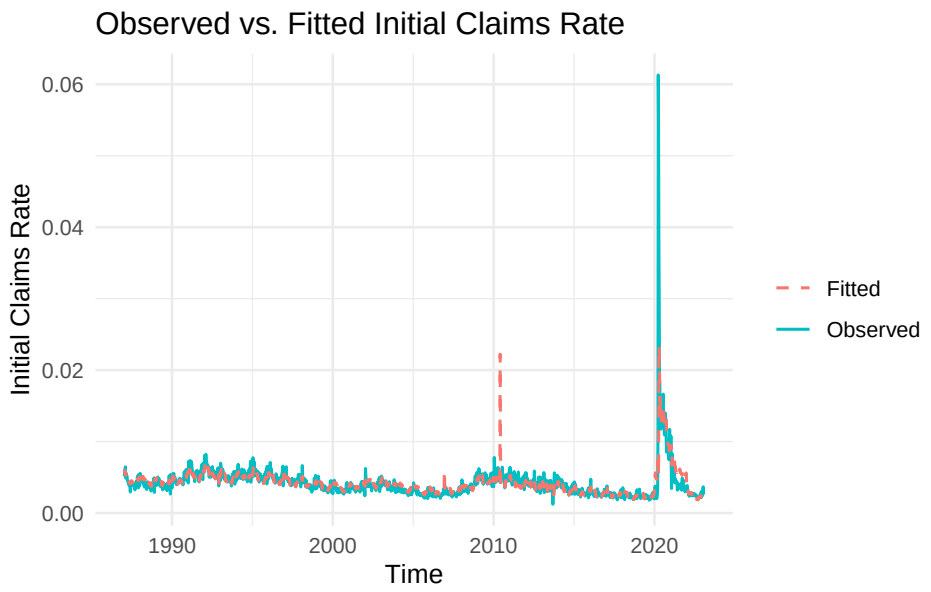

Figure 8: Observed and fitted values of the initial unemployment insurance claims rate from the selected model.

As shown in Figure 8, the selected model tracks the observed initial claims rate closely during typical periods and captures the size and persistence of major shocks, particularly during the COVID-19 period.

Discussion

This paper examines the dynamics of weekly unemployment insurance activity in California using nearly four decades of data. By moving from a simple contemporaneous regression to progressively richer time dependent models, the analysis aims to identify how labor market conditions propagate through the unemployment insurance system over time.

Key Findings and Implications

The results demonstrate that unemployment insurance claims exhibit strong persistence and cannot be adequately explained using same-week labor market indicators alone. Lagged values of initial claims play a central role, indicating that UI activity evolves gradually rather than responding instantaneously to economic conditions. This finding highlights the importance of accounting for temporal dependence when modeling unemployment-related outcomes.

In addition, the analysis shows that major economic shocks introduce sharp structural changes in UI dynamics. The COVID-19 period, in particular, is qualitatively different from prior recessions, producing unprecedented spikes in insured unemployment that dominate the upper tail of the data. These results suggest that standard recession-based modeling assumptions may be insufficient during extreme events and that explicit structural break indicators are necessary to capture such disruptions.

From a policy perspective, the findings imply that unemployment insurance systems respond with delay and persistence, meaning that elevated claims can continue even after broader economic conditions begin to improve. This has important implications for UI administration, forecasting workloads, and fiscal planning during and after economic downturns.

Limitations and Recomendations

Despite substantial improvements over the baseline specification, several limitations remain. Residual diagnostics indicate that some degree of heteroskedasticity and non-normality persists, particularly during periods of extreme economic stress such as the COVID-19 pandemic. While structural break indicators mitigate these effects, they cannot fully eliminate the impact of unprecedented shocks in a linear framework.

Additionally, the assumption of independent and identically distributed errors remains an approximation, as weekly unemployment data inherently exhibit complex temporal dependence. More formal time-series models, such as ARIMA or state-space frameworks, may further improve residual behavior but fall outside the scope of this analysis.

Future extensions could explore nonlinear dynamics, regime-switching models, or distributional approaches such as Gamma regression to better capture the positive and skewed nature of unemployment claim rates. Nonetheless, the selected model provides a transparent and statistically sound foundation for understanding how unemployment insurance activity responds to evolving labor market conditions over time.

Future Extensions

Hence, the possible future research could extend this analysis by adopting fully specified time-series frameworks, such as ARIMA or state-space models, which explicitly model serial correlation and evolving variance structures. Regime-switching or threshold models could further capture differences between normal economic conditions and crisis periods. Additional extensions could incorporate real-time policy changes, eligibility rules, or demographic heterogeneity to better understand how different groups interact with the unemployment insurance system.

Software

All analyses were conducted in R with dplyr for wrangling, ggplot2 + ggpubr for graphics, broom for model augmentations.

References

Federal Reserve History. 2013. "The Great Recession and Its Aftermath." 2013. https://www.federalreservehistory.org/essays/great-recession-and-its-aftermath.

Gardner, Jennifer M. 1994. "The 1990–91 Recession: How Bad Was the Labor Market?" *Monthly Labor Review*. https://www.bls.gov/mlr/1994/06/art1full.pdf.

Neal, Michael, and Laurie Goodman. 2021. "Understanding the Differences Between the COVID-19 Recession and the Great Recession Can Help Policymakers Implement Successful Loss Mitigation." May 2021. https://www.urban.org/urban-wire/understanding-differences-between-covid-19-recession-and-great-recession-can-help-policymakers-implement-successful-loss-mitigation.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

U.S. Bureau of Labor Statistics. 2012. "The Recession of 2007–2009: BLS Spotlight on Statistics." 2012. https://www.bls.gov/spotlight/2012/recession/pdf/recession_bls_spotlight.pdf.

Walters, Dan. 2022. "California Finally Regains Jobs Lost in COVID-19 Recession." 2022. https://calmatters.org/commentary/2022/08/california-finally-regains-jobs-lost-in-covid-19-recession/.

YCharts. 2025. "California Labor Force." 2025. https://ycharts.com/indicators/california_labor_force.