Effects of Remote Work on Wages
A Bayesian Perspective

Aleksandr Dudakov

2025-05-16

Abstract

This study examines the causal impact of remote work on hourly wages in the post-pandemic U.S. labor market using a two-stage Bayesian propensity score analysis (BPSA). Using microdata from the January 2025 Current Population Survey Outgoing Rotation Group, individuals are stratified into quintiles based on their estimated propensity to telework. In the design stage, Bayesian logistic regression with Metropolis-Hastings sampling is used to derive propensity scores while fully accounting for estimation uncertainty. In the subsequent analysis stage, stratified Bayesian linear regression with Gibbs sampling models log wage outcomes, and Rubin's rules are applied to integrate both within-design and between-design variability. The results indicate a statistically robust and substantively meaningful wage premium for teleworkers, with an average treatment effect (ATE) of 0.164 on the log scale - equivalent to approximately a 17.8% increase in hourly wages. By incorporating design uncertainty into the estimation process, the Bayesian framework provides deeper, more reliable inference than traditional frequentist approaches. These findings advance our understanding of the economic effects of remote work and have important implications for policymakers, employers, and workers in the changing context of labor market practices.

University of Milan
Department of Economics, Management and Quantitative Methods (DEMM)
Bayesian Analysis
Prof. Luca Rossini

UNIVERSITÀ
DEGLI STUDI
DI MILANO

DEMM

1 Introduction

Wages are shaped by a variety of factors, including human capital, job characteristics, and market conditions. In recent years, *remote work* has emerged as a significant factor influencing employment practices, particularly following the COVID-19 pandemic. The rapid expansion of telework in 2020-2021 has changed perceptions of remote work, shifting its role from a niche benefit to a mainstream work arrangement. As a result, accurately identifying the causal effect of remote work on wages has become critical for policymakers, employers, and workers. However, this task is complicated by self-selection and endogeneity, as workers who choose to work remotely often differ systematically from their on-site counterparts.

Several empirical studies have addressed these challenges. Oettinger (2011) documents a substantial decline in the wage penalty for home-based employment in the U.S. between 1980 and 2000, attributing this shift primarily to reduced employer costs associated with advances in information technology. More recently, Pabilonia and Vernon (2025) highlights a positive wage premium for remote work, which increased significantly during the COVID-19 pandemic, along with higher wage growth relative to on-site employment. In addition, Mas and Pallais (2017) illustrates heterogeneous worker valuations of remote arrangements, with some individuals willing to sacrifice wages for greater flexibility.

Methodologically, propensity score analysis (PSA) is often used in observational studies to mitigate selection bias by adjusting for covariates between treated and control groups. However, conventional frequentist PSA implementations typically treat estimated propensity scores as known quantities, neglecting estimation uncertainty and thus potentially overstating the precision of causal effects. *Bayesian Propensity Score Analysis (BPSA)* addresses this critical limitation by explicitly incorporating uncertainty from the design stage into the final estimation, producing posterior distributions of treatment effects that more accurately reflect true uncertainty in inference. Kaplan and Chen (2012) introduced a two-stage BPSA procedure that first estimates propensity scores under Bayesian logistic regression and then draws multiple imputations of these scores into the second-stage outcome model. This approach avoids feedback from the outcome to the propensity model while propagating uncertainty through both stages. More recently, Liao and Zigler (2020) provided a systematic way to incorporate variability across different propensity score implementations (matching, weighting, stratification), demonstrating that uncertainty at the design stage can significantly affect final treatment effect estimates.

Building on these methodological advances, this study applies a stratification-based Bayesian propensity score framework following Liao and Zigler (2020) to rigorously assess the causal impact of remote work on wages. Using representative individual-level microdata from the January 2025 Current Population Survey (CPS) Outgoing Rotation Group (ORG), the analysis captures a

stabilized post-pandemic labor market environment in which remote work has become widely institutionalized. Unlike previous studies, which were limited by the volatility immediately following the pandemic, this research provides updated findings that reflect current labor market realities. The comprehensive coverage of the CPS ORG dataset allows for the estimation of an Average Treatment Effect (ATE) that is representative of the entire U.S. labor force, rather than focusing on specific subpopulations.

The Bayesian approach strengthens inference by fully integrating the uncertainty from propensity score estimation, generating posterior distributions that reflect the variability of the stratification design. This avoids the overstated precision common in frequentist frameworks and yields more reliable credible intervals.

2 Data

The Current Population Survey (CPS) is a monthly survey conducted by the U.S. Census Bureau and the Bureau of Labor Statistics (BLS) that provides comprehensive information on the characteristics of the U.S. labor force. The Outgoing Rotation Group (ORG) component specifically collects detailed earnings data, making it particularly useful for wage analysis.

This analysis uses CPS ORG data from January 2025, focusing on employed individuals aged 16 and older. The dataset was obtained using the `epiextractr` package in `R`, which facilitates the download and loading of Economic Policy Institute (EPI) microdata extracts. EPI CPS ORG extracts include only individuals with positive sample weights who are in the outgoing rotation months, ensuring data reliability and consistency.

Several preprocessing steps were performed: age was originally limited to "80+" with 68 observations that were recoded to age 80. Similarly, the "99+" category for weekly hours worked, with 11 observations, was recoded to 99. Given their small number, these recodings should have minimal impact on the analysis.

Initial data inspection identified missing values for critical variables: wage (11.3%), telework (3.4%), hours (3.4%), union membership (11.0%), and metropolitan status (0.8%). To address this, hot deck imputation-a method recommended by the U.S. Census Bureau-was used to replace missing values with responses from similar respondents matched on age, race, gender, and education. After imputation, the overall missing rate dropped to 1.8%, and these remaining missing cases were removed. Sensitivity analyses conducted using only complete case data yielded practically similar results (not shown), confirming the robustness of the imputation process.

The final dataset contains 11,574 observations and includes the variables listed in Table 1, with all categorical levels explicitly listed. These variables were chosen because they capture important sociodemographic, economic, and occupational characteristics that influence both an

individual's likelihood of teleworking and their wage outcomes. Including factors such as age, education, and industry ensures that the propensity score model adjusts for potential confounding and achieves a balanced comparison between teleworkers and non-teleworkers.

Exploratory Data Analysis A log transformation was applied to the `wage` variable, which significantly reduced its original right-skewness and stabilized its variance. After transformation, the distribution of `log_wage` approximates normality, with only slight deviations at the tails. Correlation analysis reveals modest positive correlations of `log_wage` with `age` ($\rho \approx 0.19$) and `telework` ($\rho \approx 0.18$), and a modest negative association with `female` ($\rho \approx -0.14$). These results underscore the importance of controlling for demographic and occupational covariates in subsequent modeling.

In addition, the median log wage is higher for teleworkers than for non-teleworkers, although the substantial overlap between the groups highlights potential selection bias. These findings justify the use of a Bayesian propensity score framework that effectively accounts for observed covariate imbalances and explicitly accounts for the uncertainty associated with observational data.

Table 1: Description of Variables in the Final Dataset

Variable	**Description**
`age`	Age of the respondent in years (16–80).
`sex`	Sex of the respondent: *Male*, *Female*.
`married`	Marital status: *Not married*, *Married*.
`children`	Presence and age group of own children under 18 years: *No own children*, *All 0–2*, *All 3–5*, *All 6–13*, *All 14–17*, and combinations thereof.
`citizen`	Citizenship status: *Not a US citizen*, *US citizen*.
`race`	Race/ethnicity: *White*, *Black*, *Hispanic*, *Asian*, *Other*.
`metropolitan`	Metropolitan status: *Nonmetropolitan*, *Metropolitan*.
`state`	Census code for the state of residence (e.g., *NY*, *CA*)
`education`	Level of education: *Less than high school*, *High school*, *Some college*, *College*, *Advanced*.
`job_class`	Class of worker: *Government - Federal*, *Government - State*, *Government - Local*, *Private, for profit*, *Private, nonprofit*, *Self-employed, incorporated*, *Self-employed, unincorporated*.
`industry`	Major industry of employment: *Agriculture, forestry, fishing, and hunting*, *Mining*, *Construction*, *Manufacturing*, *Wholesale and retail trade*, *Transportation and utilities*, *Information*, *Financial activities*, *Professional and business services*, *Educational and health services*, *Leisure and hospitality*, *Other services*, *Public administration*.
`hours`	Total number of hours worked last week.
`union`	Union membership status.
`telework`	Teleworked or worked from home for pay at any time last week.
`wage`	Hourly wage in dollars, including overtime, tips, and commissions.

3 Methodology

3.1 Causal Estimation Framework & Propensity Score Stratification

In this subsection, I formalize the causal estimand and identification strategy for assessing the impact of telework on wages, following Cunningham, 2021.

Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes (wages) for individual i under the treatment (telework) and control (on-site work), respectively. Under SUTVA, the observed outcome is given by $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$, where $D_i \in \{0, 1\}$ denotes treatment status. The individual causal effect is $\delta_i = Y_i(1) - Y_i(0)$, and the Average Treatment Effect (ATE) is defined as $\text{ATE} = \mathbb{E}[\delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$.

1. **Conditional Ignorability:** $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$, where X_i is the vector of socio-demographic covariates defined in Table 1 (all variables, except `telework` and `wage`).

2. **Common Support:** $0 < \Pr(D_i = 1 \mid X_i) < 1 \quad$ for all i.

These assumptions ensure that, conditional on X_i, treatment is effectively randomized and that each individual has a non-zero chance of receiving either treatment.

The propensity score is defined as $e(X_i) = \Pr(D_i = 1 \mid X_i)$. By Rosenbaum and Rubin, 1983, $e(X_i)$ is a balancing score such that, conditional on $e(X_i)$, the distribution of X_i is the same for treated and control units.

To adjust for residual confounding, I stratify the sample into S strata (with $S = 5$ for quintile stratification) based on the quantiles of the estimated propensity score $e(X_i)$. Let $\nu_i \in \{1, \ldots, S\}$ denote the stratum assignment for observation i. I model the outcome using a stratified linear regression:

$$E(log(Y_i)) = \beta_0 + \beta_1 D_i + \sum_{s=2}^{S} \Big[\beta_{2s}\, 1(\nu_i = s) + \beta_{3s}\, D_i\, 1(\nu_i = s) \Big], \tag{1}$$

where β_1 is the treatment effect in the reference stratum ($s = 1$), and β_{2s} and β_{3s} capture, respectively, the baseline shift and the differential treatment effect in stratum $s \geq 2$.

Define $P_{st} = \frac{\sum_{i=1}^{n} 1(\nu_i = s)\, 1(D_i = t)}{\sum_{i=1}^{n} 1(D_i = t)}$ as the proportion of units in treatment group t assigned to stratum s. Then, the aggregated Average Treatment Effect (ATE) is given by:

$$\text{ATE} = \beta_1 + \sum_{s=2}^{S} P_{s1}\, \beta_{3s}. \tag{2}$$

3.2 Design Stage: Bayesian Propensity Score Estimation via Metropolis-Hastings

In the design stage, I estimate the propensity score $e(X_i)$ using a Bayesian logistic regression model. Formally, the likelihood for the binary treatment indicator D_i follows a Bernoulli distri-

bution:

$$D_i \sim \text{Bernoulli}(\pi_i), \quad \text{with} \quad \pi_i = \frac{\exp(X_i^\top \theta)}{1 + \exp(X_i^\top \theta)},$$

where the vector θ denotes the parameters in the logistic regression that capture the effect of the covariates in X_i on the probability of treatment assignment. Hence, the joint likelihood across observations is $L(\theta; X, D) = \prod_{i=1}^{n} \left[\frac{\exp(X_i^\top \theta)}{1+\exp(X_i^\top \theta)}\right]^{D_i} \left[\frac{1}{1+\exp(X_i^\top \theta)}\right]^{1-D_i}$.

Given limited prior knowledge, I impose a weakly informative multivariate normal prior on θ:

$$\theta \sim N(b_0, B_0^{-1}), \quad \text{with } b_0 = 0,\ B_0 = 0.1 \cdot I.$$

This prior regularizes the estimates, stabilizes the inference, and ensures that the posterior estimates are primarily data-driven.

By Bayes' theorem, the posterior distribution of parameters θ is

$$\pi(\theta \mid X, D) \propto L(\theta; X, D) \cdot \pi(\theta).$$

I approximate this posterior distribution using a Metropolis-Hastings random walk sampler because the posterior distribution is not available in closed form, and this approach allows us to efficiently sample from the target distribution despite the non-conjugacy of the logistic model. At iteration j:

1. Initialize $\theta^{(0)}$ at maximum likelihood estimates.
2. Propose a candidate θ^* from a symmetric normal random walk: $\theta^* \sim N(\theta^{(j-1)}, \Sigma)$, where the proposal covariance matrix Σ is tuned to achieve an optimal acceptance rate (20–30%).
3. Accept candidate θ^* with probability $\alpha = \min\left\{1, \frac{\pi(\theta^*|X,D)}{\pi(\theta^{(j-1)}|X,D)}\right\}$. Otherwise, retain $\theta^{(j-1)}$.
4. Iterate for 130,000 steps per chain, discarding the initial 5,000 iterations as burn-in and thinning the remaining draws by retaining every 500th iteration.

I run four parallel chains, resulting in 1,000 posterior draws of θ. For each posterior draw, the propensity score is computed as: $e(X_i) = \frac{1}{1+\exp(-X_i^\top \theta)}$. This Bayesian estimation fully incorporates posterior uncertainty from the design stage into subsequent causal analyses.

3.3 Analysis Stage: Bayesian Stratification via Gibbs Sampling

In the analysis stage, I estimate the stratified outcome model (1) via a Bayesian linear regression conditioned on the propensity score strata obtained from the design stage:

$$\log(Y_i) \;=\; \hat{\beta}_0 \;+\; \hat{\beta}_1 D_i \;+\; \sum_{s=2}^{S} \Big[\hat{\beta}_{2s}\, 1(\hat{\nu}_i = s) \;+\; \hat{\beta}_{3s}\, D_i\, 1(\hat{\nu}_i = s)\Big] \;+\; \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. This way I can estimate the Average Treatment Effect (ATE) using (2).

Conditioned on $\hat{\nu}_i$, each observation i follows a Gaussian likelihood:

$$\log(Y_i) \,\big|\, (D_i, \hat{\nu}_i, \boldsymbol{\beta}, \sigma^2) \;\sim\; N\Big(\hat{\beta}_0 \;+\; \hat{\beta}_1 D_i \;+\; \sum_{s=2}^{S}\big[\hat{\beta}_{2s}\, 1(\hat{\nu}_i = s) \;+\; \hat{\beta}_{3s}\, D_i\, 1(\hat{\nu}_i = s)\big],\; \sigma^2\Big).$$

Hence, across the sample of size n, the joint likelihood is

$$L(\boldsymbol{\beta}, \sigma^2;\, Y, D, \hat{\nu}) \;=\; (2\pi\sigma^2)^{-\frac{n}{2}} \exp\Big(-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n}\big[\log(Y_i) - \hat{\beta}_0 - \hat{\beta}_1 D_i - \sum_{s=2}^{S}\big(\hat{\beta}_{2s}\, 1(\hat{\nu}_i = s) + \hat{\beta}_{3s}\, D_i\, 1(\hat{\nu}_i = s)\big)\big]^2\Big).$$

I assign an *improper uniform prior* to the regression coefficients $\boldsymbol{\beta}$, reflecting minimal prior information:

$$p(\boldsymbol{\beta}) \;\propto\; 1.$$

For the error variance, I specify a standard conjugate inverse-gamma prior with weak parameters:

$$\sigma^2 \;\sim\; \text{Inv-Gamma}\Big(\tfrac{c_0}{2},\, \tfrac{d_0}{2}\Big), \quad c_0 = 0.001, \quad d_0 = 0.001,$$

ensuring that posterior inferences remain predominantly data-driven.

Because the model is conditionally conjugate, I perform Gibbs sampling by alternating between two full conditional draws:

1. **Regression coefficients $\boldsymbol{\beta}$.** Conditioned on σ^2, the conditional posterior for $\boldsymbol{\beta}$ is multivariate normal:

$$\boldsymbol{\beta} \,\big|\, (\sigma^2, Y, D, \hat{\nu}) \;\sim\; N\Big(\big(X^\top X\big)^{-1} X^\top \log(Y),\; \sigma^2 \big(X^\top X\big)^{-1}\Big),$$

   where X is the design matrix incorporating an intercept, the treatment indicator D_i, stratum indicators $1(\hat{\nu}_i = s)$, and all relevant interactions.

2. **Error variance σ^2.** Conditioned on $\boldsymbol{\beta}$, the variance parameter follows an inverse-gamma posterior:

$$\sigma^2 \;\big|\; (\boldsymbol{\beta}, Y, D, \hat{\nu}) \;\sim\; \text{Inv-Gamma}\Big(\tfrac{n+c_0}{2},\; \tfrac{SSR + d_0}{2}\Big),$$

   where $SSR \;=\; \sum_{i=1}^{n}\Big[\log(Y_i) \;-\; \hat{\beta}_0 \;-\; \hat{\beta}_1 D_i \;-\; \sum_{s=2}^{S}\big(\hat{\beta}_{2s}\, 1(\hat{\nu}_i = s) \;+\; \hat{\beta}_{3s}\, D_i\, 1(\hat{\nu}_i = s)\big)\Big]^2$.

In each analysis, I discard the first 500 draws as burn-in and collect a total of 1000 samples, thinning the remaining draws by retaining every 5th iteration, resulting in 200 draws per run. This sampling is repeated for each posterior draw of the propensity scores (i.e., each design stage draw).

From each chain, I extract posterior samples of $(\boldsymbol{\beta}, \sigma^2)$ and derive draws of the ATE by combining $\hat{\beta}_1$ with interaction effects across strata. I then pool these samples across all propensity score draws, applying Rubin's rules to account for both *within-design* and *between-design* variability. In addition, I estimate the posterior mean of the ATE, as a point estimate; 95% equi-tailed credible intervals, obtained via the 2.5th and 97.5th posterior percentiles; and highest density intervals (HDIs) for a sharper focus on the most probable parameter regions.

3.4 Convergence and Diagnostic Analysis (CODA)

To ensure reliable Bayesian inference, I evaluate MCMC convergence and mixing using the following diagnostics:

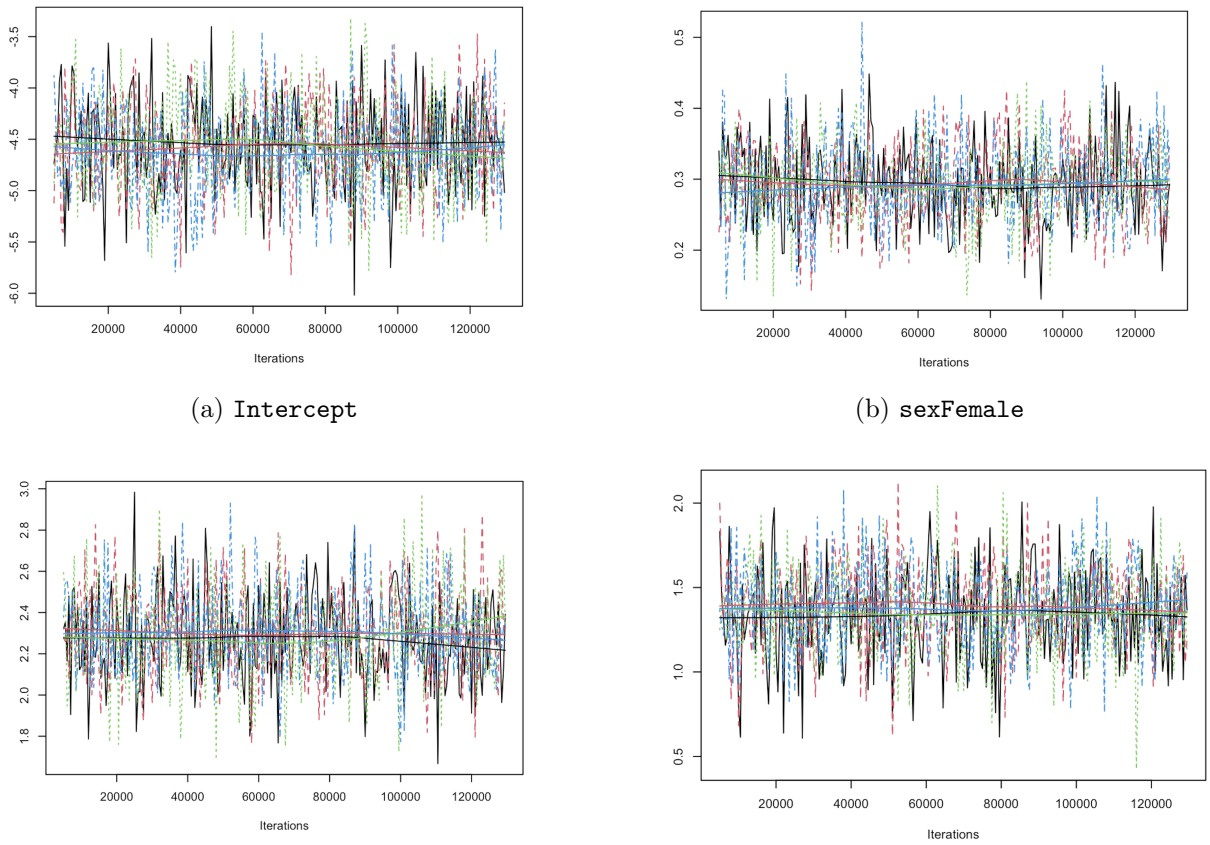
- **Traceplots:** Visual inspection of parameter draws across iterations. Converged chains show stationarity and thorough posterior exploration. Persistent trends or stagnation indicate inadequate mixing, requiring tuning or additional iterations.
- **Autocorrelation Analysis:** Autocorrelation functions (ACFs) at selected lags (e.g., 0, 1, 5, 10, 100) measure the dependence between successive draws. Rapid decay indicates efficient mixing, while high autocorrelation at large lags indicates slow exploration and may require increased thinning or longer runs.
- **Gelman-Rubin Diagnostic:** The Gelman-Rubin statistic ($\hat{R}$) compares the between-chain and within-chain variance across multiple chains. Values close to 1 (typically less than 1.1) indicate convergence; higher values indicate persistent between-chain discrepancies that require additional iterations.

For the design stage, which uses a Metropolis-Hastings algorithm for logistic regression, four parallel chains were monitored using traceplots, ACFs, and the Gelman-Rubin diagnostic to address the challenge of non-conjugacy. In contrast, the analysis stage uses Gibbs sampling with conditional conjugacy, where visual inspection of traceplots and ACFs was sufficient to confirm rapid mixing and convergence.

4 Empirical Results

In this section, I present the estimated causal effects of telework on wages using the two-stage Bayesian Propensity Score Analysis (BPSA) described above. First, I summarize the MCMC diagnostics and posterior estimates for the propensity score model at the design stage. Second, I describe the results of the Bayesian outcome model under quintile-based propensity score stratification, paying particular attention to the posterior distributions of the Average Treatment Effect

(ATE). Third, I provide numerical and graphical plots of the posterior draws to illustrate the robustness of the results.



(a) `Intercept` (b) `sexFemale`

(c) `educationCollege` (d) `industryInformation`

Figure 1: Metropolis–Hastings trace plots for selected propensity score coefficients. Each panel shows overlapping trajectories from four parallel MCMC chains, indicating stable convergence and good mixing.

4.1 Design Stage: Bayesian Propensity Score Estimation via Metropolis-Hastings

Bayesian modeling in this study is based on the R package `MCMCpack`. Estimation of the propensity score uses Metropolis-Hastings sampling for 125,000 iterations per chain (plus 5,000 burn-in), thinned at an interval of 500. Four parallel chains were run to facilitate convergence diagnostics. Visual inspection of the parameter traceplots (selected traceplots are shown in Figure 1) showed no prolonged trends or block-like patterns, and the density plots were unimodal and smooth, suggesting stable posterior exploration. The autocorrelation functions (ACFs) across successive draws decayed rapidly, indicating that thinning at 500 was sufficient to mitigate within-chain dependence.

A formal Gelman-Rubin diagnostic (Table 2) yielded potential scale reduction factors $\hat{R}$ close to 1.00 for all slope parameters, with an upper bound of about 1.02 for the slowest mixing parameters. The overall multivariate potential scale reduction factor was 1.07, well below the

commonly accepted threshold of 1.1. These results confirm adequate chain convergence and mixing at the design stage.

Table 2: Excerpt of Gelman–Rubin Diagnostic for Selected Propensity Score Coefficients

Parameter	**Point est.**	**Upper C.I.**	**Multivariate $\hat{R}$**
`(Intercept)`	1.003	1.013	≈ 1.07
`sexFemale`	1.000	1.000	
`educationCollege`	1.003	1.011	
`industryInformation`	1.005	1.017	
⋮	⋮	⋮	

Note: Only four representative parameters shown here. All model coefficients had $\hat{R} < 1.03$. The multivariate $\hat{R} = 1.07$ is below the 1.1 threshold, indicating convergence.

Table 3 (excerpted) reports the posterior means and 95% credible intervals for a subset of the logistic regression coefficients. For example, `sexFemale` is positively associated with an increased probability of teleworking, controlling for other covariates. Higher levels of education (`college`, `advanced`) also show strong positive associations with teleworking, consistent with the well-documented pattern that teleworking is more common in professional, IT, and knowledge-based occupations that require higher levels of human capital. In contrast, residence in certain states (e.g., `MS`, `AL`) is negatively associated with telework, suggesting possible regional differences in telework adoption.

Posterior draws of each individual's propensity score were generated via the logistic inverse link: $e(X_i) = \frac{1}{1+\exp(-X_i^\top \theta)}$), where θ is sampled from its converged posterior distribution (1,000 draws remaining after burn-in and thinning). This procedure yielded a $n \times 1000$ matrix of propensity scores, with $n = 11,574$. Inspection of the distribution of posterior propensity scores (not shown) revealed broad coverage across the [0,1] interval, although heavier mass was observed below 0.3, consistent with the relatively small proportion of teleworkers ($\approx$ 24%) in the sample.

Table 3: Selected Posterior Summaries for the Bayesian Logistic Propensity Score Model

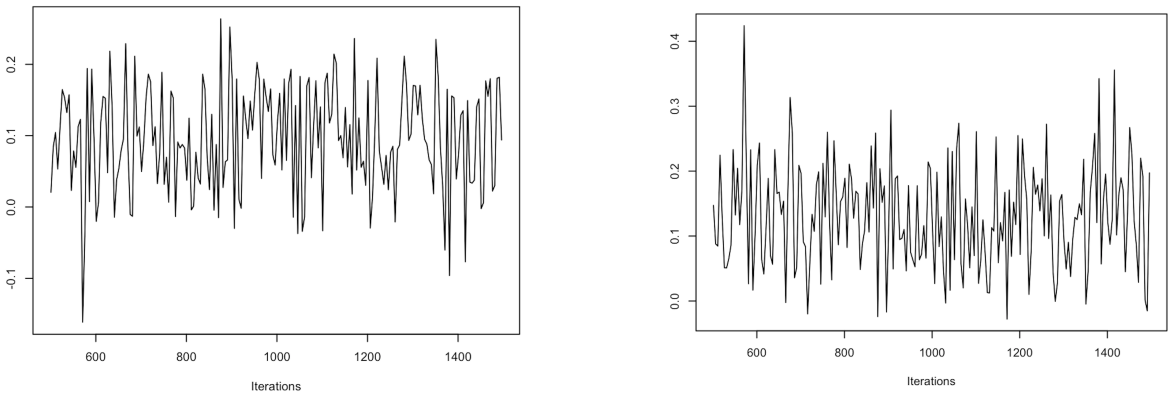
Coefficient	**Mean**	**SD**	**2.5%**	**97.5%**
`(Intercept)`	-4.579	0.447	-5.457	-3.720
`age`	0.0069	0.0021	0.0025	0.0107
`sexFemale`	0.2934	0.0560	0.1848	0.4074
`marriedMarried`	0.1026	0.0609	-0.0200	0.2183
`educationCollege`	2.2926	0.2153	1.8793	2.7374
`hours`	0.0110	0.0022	0.0067	0.0157
⋮	⋮	⋮	⋮	⋮

Note: Posterior means, standard deviations (SD), and 95% credible intervals (2.5% and 97.5%) for an illustrative subset of the logistic regression coefficients.

4.2 Analysis Stage: Bayesian Stratification via Gibbs Sampling

For each of the 1,000 posterior draws of θ from the design stage, I computed individual-level propensity scores and stratified the sample into five strata (quintiles).

The outcome model took the form $\log(Y_i) = \beta_0 + \beta_1 D_i + \sum_{s=2}^{S} \big[\beta_{2s}\, 1(\nu_i = s) + \beta_{3s}\, D_i\, 1(\nu_i = s)\big] + \varepsilon_i$, where $D_i \in \{0, 1\}$ is the treatment (telework) indicator, $\nu_i \in \{1, \ldots, 5\}$ is the stratum assignment, and $\varepsilon_i \sim N(0, \sigma^2)$. Each of the five strata thus allowed for a distinct baseline mean and a distinct incremental telework effect (through the interactions). All parameters were assigned conditionally conjugate priors, allowing for simple Gibbs sampling. Within each propensity score draw, the outcome model was sampled for 200 draws (after burn-in and thinning). Figure 2 shows representative traceplots for several analysis stage coefficients; again, the chains mix well, and no major autocorrelation was evident at lags beyond 5 or 10.



(a) `treatment` (b) `treatment:strata3`

Figure 2: Representative Gibbs sampling traceplots for selected analysis stage coefficients. The draws appear well-mixed with no drifting or non-stationary behavior.

4.3 Average Treatment Effect (ATE) Estimates

Combining the stratified regression coefficients as in equation (2) yielded stratum-specific partial effects, which were then aggregated using the relative stratum sizes. Repeating this procedure over the entire set of propensity score draws yielded $1,000 \times 200 = 200\,000$ draws of the ATE. Figure 3 shows the histogram of these pooled draws, which is approximately normal with a clearly positive mean.

Formally applying Rubin's rules, the total variance of the ATE estimates decomposes into *within-design* (MCMC sampling within each propensity score draw) and *between-design* (variability across the different propensity score draws). The results, reported in Table 4, indicate that the between-design component is relatively modest, meaning that differences in estimated propensity scores across draws do not significantly alter the estimated effect.

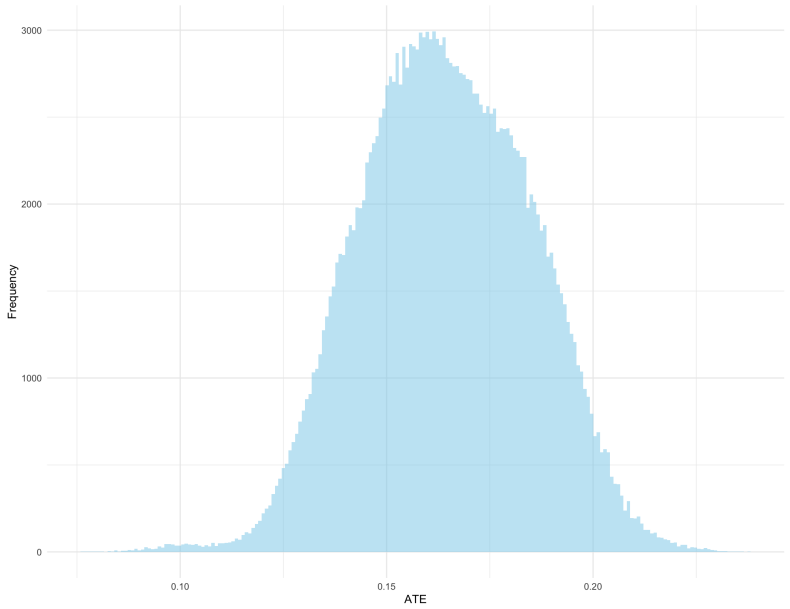


Figure 3: Posterior distribution of the Average Treatment Effect (ATE) on log-wages from the Bayesian stratified model, combining across all propensity score draws.

Table 4: Final Estimates of the ATE on Log-Wages

Posterior Quantity	**Value**
Within-design variance, $\bar{V}_{\text{within}}$	0.000364
Between-design variance, V_{between}	0.000056
Total variance, V_{total}	0.000420
Posterior Mean of ATE, $\bar{\delta}$	0.164
95% ETI (Quantile-based)	[0.126, 0.203]
95% HDI	[0.127, 0.203]
Wald-type 95% CI	[0.124, 0.204]

Notes: ETI denotes the 2.5% and 97.5% equi-tailed credible interval, while HDI is the 95% highest density interval. The Wald-type interval applies a normal approximation with the total variance from Rubin's rules.

The posterior mean of the ATE on the log of wages is 0.164, with narrow 95% credible intervals: $[0.126, 0.203]$ by either equal-tailed or highest density criteria. Since the outcome model uses $\log(Y_i)$, we can exponentiate the effect to interpret it as a proportional wage difference: $\exp(0.164) \approx 1.178$, implying that telework increases hourly wages by about 17.8% on average, ceteris paribus. This effect is both statistically credible (near-zero posterior mass below zero) and substantively meaningful, consistent with previous findings (Pabilonia and Vernon, 2025) that teleworkers receive wage premiums.

Although the estimates suggest a robust positive premium, it is important to recall that these results are based on the assumption of unconfoundedness, i.e., no substantial omitted variable bias after conditioning on the rich set of CPS covariates. Potential unmeasured confounders related to individual preferences, job tasks, and firm-level policies may still remain. Nevertheless, within the framework of fully observed covariates, the Bayesian propensity score approach

incorporates uncertainty at the design stage and suggests a telework wage premium with high posterior confidence.

Additional sensitivity analyses were performed to assess the robustness of the causal estimates. Excluding imputed cases (i.e., using only complete case data) yielded a treatment effect that was approximately the same as in the main analysis. Similarly, varying the number of propensity score strata (using 3 or 10 strata) resulted in nearly unchanged estimates.

Overall, the above diagnostics, posterior summaries, and robustness checks collectively confirm a credible, positive telework effect on wages in the U.S. labor market as of January 2025.

5 Conclusion

This study rigorously examines the causal effect of remote work on wages by implementing a two-stage Bayesian Propensity Score Analysis. Using a large, representative CPS ORG sample from January 2025 and stratifying individuals into quintiles of their telework propensity, it estimated an Average Treatment Effect (ATE) of approximately 17.8% wage premium for teleworkers, with narrow credible intervals. By explicitly incorporating uncertainty from the design stage into the final outcome model, the analysis avoids exaggerated precision and more accurately represents the underlying variability in the data. These findings underscore the persistent importance of telework in shaping wage outcomes in the post-pandemic period and are consistent with emerging research linking remote work to higher compensation.

The findings invite deeper reflection on the channels through which telework may provide economic benefits. Remote work is thought to increase productivity by eliminating long commutes and reducing workplace distractions. These efficiency gains are likely to increase individual output while reducing non-wage costs for employers - savings that may be partially passed on to workers in the form of higher wages. In addition, cost savings on the employer side, such as reduced spending on office space and related costs, may further contribute to wage increases. Firms, under competitive pressure to attract and retain top talent, may share these savings with employees, thereby reinforcing the observed wage premium.

Despite the methodological rigor, several limitations remain. The analysis relies on unconfoundedness, which assumes that no critical unmeasured variables systematically distinguish teleworkers from on-site workers. In addition, the data cover a single time period, which prevents direct examination of dynamic trends. Finally, although missing data were handled using hot deck imputation, residual bias could persist if imputation models or assumptions about missingness are incomplete.

Future research could use longitudinal or repeated cross-sectional data to examine how the wage effects of telework evolve over time. The incorporation of additional covariates, such as

detailed job-task measures or firm-level characteristics, could further mitigate omitted variable bias and help identify the mechanisms underlying the telework premium. Finally, incorporating hierarchical models or sector-specific analyses would improve our understanding of how the benefits and challenges of telework vary by region, industry, and job type, providing a more nuanced perspective on this growing work arrangement.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Cunningham, S. (2021). *Causal inference: The mixtape.* Yale university press.

Hoff, P. D. (2009). *A first course in bayesian statistical methods* (Vol. 580). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r.* Springer.

Kaplan, D., & Chen, J. (2012). A two-step bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, *77*(3), 581–609.

Liao, S. X., & Zigler, C. M. (2020). Uncertainty in the design stage of two-stage bayesian propensity score analysis. *Statistics in medicine*, *39*(17), 2265–2290.

Mas, A., & Pallais, A. (2017). Valuing alternative work arrangements. *American Economic Review*, *107*(12), 3722–3759.

Oettinger, G. S. (2011). The incidence and wage consequences of home-based work in the united states, 1980–2000. *Journal of Human Resources*, *46*(2), 237–260.

Pabilonia, S. W., & Vernon, V. (2025). Remote work, wages, and hours worked in the united states. *Journal of Population Economics*, *38*(1), 1–49.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rossini, L. (2025). *Bayesian analysis: Course notes.* University of Milan.