

Remote vs. Office Work: Understanding the Key Determinants in Australia using Bayesian logistic regression



Abstract

In the wake of the COVID-19 pandemic, the world witnessed a significant shift toward remote work as a necessary response to curbing the virus's spread. This transition raised a multitude of challenges, from juggling work with childcare to adapting to virtual meetings. However, as time passed, many of these challenges were addressed. In a post-pandemic landscape, a diverse group of Australian professionals were surveyed to explore their workplace preferences, encompassing both remote and office-based work. In the data (we consider only 2021 dataset), there were many missing values and redundant variables that have been removed. The analysis leveraged Bayesian Logistic Regression to uncover the key determinants influencing these preferences. To fit Bayesian logistic regression model Metropolis Hastings MCMC algorithm has been used. There were total 195 variables out of which 23 were important variables selected based on the Bayesian method. The results of this study highlight several noteworthy trends. Industries such as finance and insurance appear less inclined to embrace remote work, while the administrative and support sectors demonstrate a preference for office-based work. The single-family households show a tendency for the office. Moreover, the report delves into the impact of organizational support for remote work, changes in barriers, and individual motivation, shedding light on the factors that influence work preferences. This report offers valuable insights into the evolving dynamics of work preferences in a post-pandemic world, providing a nuanced understanding of the choices individuals make regarding remote or office-based work.

Contents

1. Introduction	3
2. Methodology	3
2.1 Model	3
2.2 Hyper Parameters Selection	4
2.3 Metropolis Hastings	5
3. Results	5
4. Conclusions	8
Appendix	15

Abbreviations	Definitions
SA	Somewhat agree
NAND	Neither agree nor disagree
StA	Strongly agree
StD	Strongly disagree
SD	Somewhat disagree
StP	Strongly Positive
SN	Somewhat Negative
SP	Somewhat Positive
NPNN	Neither positive nor negative
SN	Strongly Negative
SI	Somewhat improved
StaySame	Stayed about the same
NotBerr	Not a barrier for me
SW	Somewhat worsened
SigIm	Significantly improved
SigWr	Significantly worsened

1. Introduction

During the COVID-19 pandemic, people worldwide were compelled to transition to remote work in order to mitigate the spread of the virus. This shift posed various challenges, including productivity issues for those with children at home and technology-related obstacles during remote work and virtual meetings. However, as time has passed, many of these challenges have been addressed and resolved.

Now that the COVID-19 pandemic has receded, there is a growing divergence of opinions regarding preferred work environments. Some individuals favor the flexibility and convenience of remote work, while others are eager to return to the traditional office setting. This dichotomy reflects the ongoing debate and evolving work dynamics in the post-pandemic landscape.

In this report, we are analyzing data that has been gathered from a diverse group of Australian professionals employed across various industries and occupations in year 2021. In the survey total of 109 question were asked, that is, I have 109 variables. There are many research questions that need to be answered, our primary focus centers on discerning Australians' workplace preferences between remote and office-based work. I aim to uncover the key determinants that influence their choice through the application of Bayesian Logistic Regression.

2. Methodology

At the outset, the dataset contained 18,374 missing values. Some variables exhibited a pattern where certain questions were interrelated. For instance, there were two questions: 'Imagine your employer offered you the option to work remotely whenever you chose, in return for negotiating a pay cut that you both agree on. Would you be interested in negotiating?' and 'What is the biggest pay cut you would accept? Please give your answer as a percentage.' Respondents who answered 'yes' to the first question also provided a response to the second question, specifying the pay cut they were willing to accept. Conversely, those who answered 'no' to the first question did not respond to the second question. This relationship between the questions resulted in a significant number of missing values within the dataset.

To address this issue, instead of simply removing the missing values, which would substantially reduce the dataset's size, I opted to remove the variables that causes those missing values. This decision left us with only 101 missing values in the remaining data, which were subsequently removed from the dataset.

In addition to handling missing values, I identified certain questions that closely resembled the central question under investigation in our project. These similar questions were redundant for our analysis, as they essentially explored the same aspects. Consequently, I decided to remove these redundant variables from the dataset to avoid investigating the same inquiry multiple times.

I further excluded questions in the dataset that exhibited close relationships with one another, as including such questions as predictors could potentially introduce multicollinearity issues when fitting the Bayesian logistic regression model.

As the dataset is reduced, I left with 54 variables and many of them are categorical variables. Creating dummy variables for all those categories will increase variables further to 195 variables that will make analysis challenging. However, as all of the variables may not be important in our case and for that I will be doing variables selection that may reduce the number of variables.

As in the original data, variables names were very long, I made variables names short. Also to some of the questions answers (each question is variable) were also long, I made it short and used some codes for those answers, which is given in the abbreviation table above.

2.1 Model

Our question of interest is a binary variable of whether they prefer to go the work or want to work from home. For that, I need to fit Bayesian logistic regression, however, before fitting I need to decide on the prior distribution of the coefficients and derive the posterior distribution.

As the response variable is binary. This means it follows Bernoulli distribution. The likelihood function of logistic regression is

$$\mathcal{L}(Y|\beta) \propto \prod_i \left(\frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}} \right)^{y_i} \left(1 - \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}} \right)^{(1-y_i)} \quad (1)$$

where, β represents the vector of parameters, with each component associated with an explanatory variable. Since β can range from $-\infty$ to ∞ , and these parameters are assumed to be independent of each other as I do not have any prior information about the association between the parameters, I adopt a multivariate normal prior distribution for β as follows

$$f(\beta|\beta_0, \Sigma_0) \propto \exp \left(-\frac{1}{2}(\beta - \beta_0)^T \Sigma_0^{-1}(\beta - \beta_0) \right), \quad (2)$$

where β_0 is the prior vector of parameters and Σ_0 is the prior covariance matrix.

The posterior distribution will become:

$$f(\beta|Y, \mathbf{X}, \beta_0, \Sigma_0) \propto \prod_i \left(\frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}} \right)^{y_i} \left(1 - \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}} \right)^{(1-y_i)} \times \exp \left(-\frac{1}{2}(\beta - \beta_0)^T \Sigma_0^{-1}(\beta - \beta_0) \right). \quad (3)$$

For variable selection, I utilize a binary vector Z consisting of 0s and 1s, with its dimension being $p - 1$, excluding the intercept where p represents the number of explanatory variables. There is inherent uncertainty in this context, and to address it, I introduce a prior distribution for Z_j as $\gamma_j \sim \text{Bernoulli}(1/2)$. The posterior distribution is:

$$\Pr(Z_i = 1|\mathbf{X}, \gamma, \beta) = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)}, \quad (4)$$

after some algebraic manipulation equation 4 is equivalent to equation 1. Thus, the posterior of Z is

$$Z_i \sim \text{Bernoulli}\left(\frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)}\right). \quad (5)$$

This is what I use to derive the posterior odds in case of logistic regression.

2.2 Hyper Parameters Selection

The hyper parameters of the prior distributions need to be set before start of the metropolis Hastings algorithm. I set $\beta_0 = 0$ in the prior because I do not have prior knowledge about the direction (negative or positive) of the effect of explanatory variables on the response variable. I assume that the explanatory variables are not associated with the response variable by setting their effects to zero. Additionally, I chose $\Sigma_0 = 100 \cdot \mathbf{I}$ to create a weakly informative prior. This spreads the prior out and ensures that the posterior will be data-dominated. The intention is to give more weight to the observed data and have it dominate the prior information. The prior for $Z_i \sim \text{Bernoulli}(1/2)$ with probability of drawing 1 has equal to the probability of drawing 0 from the Bernoulli distribution.

2.3 Metropolis Hastings

I will be using metropolis algorithm to fit the logistic regression as follows:

1. **Initial values:** Start with a random vector, Z_i , of size $p - 1$ generated from the binomial distribution of 0s and 1s with equal probabilities. Also take initial values of the β , which I take from frequentist generalized linear model (GLM).
2. **Updating Z :** Update Z_i based on the equation 4, which is the same as equation 1 after some algebraic manipulation. Using this probability to draw $(p - 1)$ Z_i s from binomial distribution.
3. **Proposal distribution:** draw a new sample from the proposal distribution (of β). For the proposal distribution, I use a multivariate normal distribution with $\mu = \beta$ as the current value and $\Sigma = \delta \times (\Sigma_0^{-1} + \Sigma_{\text{glm}}^{-1})^{-1}$. Here, δ represents the tuning parameter, Σ_{glm} is the variance-covariance matrix from the frequentist (GLM), and Σ_0 is the prior covariance matrix. The tuning parameter value in our case is 0.02 (with acceptance 31%).
4. **Acceptance ratio:** Calculate the acceptance ratio as follows:

$$R = \min \left\{ 1, \frac{P(\text{new sample})}{P(\text{current sample})} \right\},$$

where $P(\text{current sample})$ and $P(\text{new sample})$ are the probability of accepted sample and new sample drawn from proposal distribution, receptively, using target distribution in equation 3. Note that the target distribution value will be based on the selected variables where $Z_i = 1$.

5. **Accept or Reject:** Generate random number U from the uniform distribution and check if $U < R$ than accept the sample otherwise reject it.
6. **Repeat:** Repeat 2 to 5 until convergence. In my case I use 100,000 iterations.

After running 100,000, I use first 50,000 as burnin.

3. Results

I implemented the Metropolis-Hastings algorithm to fit the model to 196 variables in our dataset. I drew 100,000 posterior draws of β from the posterior distribution given in equation 3 and did the variable selection simultaneously. I excluded the first 50,000 draws as burn-in. To check whether the posterior draws of the remaining β are converged, I examined the diagnostic plots. However, given the vast number of variables, it is not feasible to display diagnostic plots for each of them, such as trace plots. As the number of variables is very large and most of them may not be important, our focus also lies in variable selection, as discussed in Section 2. I have chosen to showcase diagnostics for a subset of important variables. The important variables that I have chosen is shown in figure 7 in the appendix, I use threshold of 0.5.

Based on the aforementioned diagnostic results shown in figure 1 & 2 of the selected variables, the posterior draws exhibit stability and a lack of significant correlations among them of some of the variables. While the trace plot of some of the variables is unstable and draws have high auto-correlation. This may be due to the fact that I do not have enough samples. I were not able to run the MCMC for larger iteration because of the high computational time. The distribution of the important variables draws is also shown in figure 3 & 4. I have identified the key variables that influence our response variable. However, it's important to note that the coefficients represent log of the odds ratios. To enhance interpretability, I am converting those estimated coefficients into odds ratios by exponentiating the estimated coefficients.

I have selected important variables, and as the log of the odds ratio is not of importance. I have converted them into odds ratios which is shown in table 2. The odds ratios and respective confidence intervals are plotted in figure 5.

In our analysis of a binary response variable, focusing on the work preferences of Australian individuals (working from home or the office), certain notable trends have emerged. Notably, those in the finance and

insurance industry are less likely to favor working from home compared to the “other services” category (other services is one of the categories of the industry variable that has been taken as a reference category), all other factors being constant. The value of the odds is 0.8446, which means individuals working in the finance and insurance industry decrease the odds ratio by 15.54% compared to the reference category, other services. Similarly, Individuals employed in administrative and support roles are more likely to choose the office (or less likely to work remotely) as their preferred work location compared to the other services, which aligns with the nature of their tasks that often require a physical presence. The odds ratio is 0.6312, which is less than one and corresponds to a decrease in the odds ratio by 36.88%. Similar patterns are observed in industries such as hospitality, where office or on-site work is favored over remote work compared to the other services, with the odds ratio 0.7658 that corresponds to a 23.42% decrease in the odds ratio. For instance, the hospitality sector predominantly requires on-site presence, as roles like restaurant cooks necessitate physical presence to prepare food.

Another determinant is related to the number of households within a residence, which can influence work preferences. Single-family households are less inclined to work remotely compared to child-rearing couples, with a 0.7349 odds ratio, causing a 26.51% decrease in the odds ratio. The reason may be that single-family households are bored at home and want to have time with their colleagues compared to the couple with little children to take care of the children as well as work.

The data collected pertains to 2021, a year when the COVID-19 pandemic was gradually receding. It is worth noting that while some organizations maintained their existing COVID-19 policies, others chose to modify them.

Initially, the work-from-home policy was implemented in response to the pandemic, with the goal of reducing the spread of COVID-19. However, as the situation improved in 2021, many people gradually returned to the office, prompting organizations to reconsider their COVID-19 policies. In this context, it is notable that some organizations chose to retain their work-from-home policy unchanged throughout 2021. Individuals employed in those organizations displayed a greater inclination to continue working remotely rather than returning to on-site work, with an odds ratio of 1.0357, causing the odds ratio to increase by 3.57% compared to people whose organizations changed their COVID-19 policy.

Individuals who expressed strong agreement with the notion that their organization actively encouraged and facilitated remote work tended to exhibit a strong preference for working remotely, with an odds ratio of 1.1950, causing a 19.5% increase in the odds ratio compared to those who somewhat agreed about the encouragement and facilitation provided by organizations. Also, those who strongly agreed about providing ease in remote work arrangements and collaboration were more inclined to opt for remote work, as evidenced by their odds ratio falling above one (1.0046) compared to those who somewhat agreed. Note that the odds ratio of arrangements and collaboration by the organizations is almost equal to one, thus, they prefer being half likely to go to the office and half likely to work remotely compared to those who somewhat agreed about the arrangements and collaboration provided by the organization. On the other hand, people who somewhat disagreed with the arrangement and collaboration provided by the organizations are less likely to prefer working remotely, with an odds ratio of 0.7501, causing a 24.99% decrease in the odds ratio compared to those who somewhat agreed. Similarly, people who strongly disagreed with the ease of collaboration and arrangements have an odds ratio of 0.5226, indicating they prefer to work in the office, causing a 47.74% decrease in the odds ratio. It is interesting to note that there is a difference between the odds ratio of people who somewhat and strongly disagree with the arrangements and collaboration of the organization, with a decrease from those who somewhat disagree to those who strongly disagree.

Some employers also prefer to work remotely. A question has been asked from the employers whether they would be impacted if they preferred to work remotely. Employers who said that they have been somewhat negatively or somewhat positively impacted prefer to work in the office or on-site compared to employers who said they were strongly positively impacted.

Sometimes various leadership, for example, managers and senior leadership in the organization, support remote work or work in the office, and that can impact the decision of whether an individual wants to work remotely or in the office. For people whose senior leadership was supportive of working remotely, they are more likely to work in the office with an odds ratio of 0.7904 compared to managers and senior leadership

who support remote work to about the same degree.

Employers who are unlikely to support remote work have employees who are less likely to work remotely compared to employers who were somewhat likely to encourage their employees to work remotely, with an odds ratio of 0.5952, causing a 40.48% decrease in the odds ratio. Employers who are very unlikely to change their support of remote working have employees who are likely to work in the office compared to employers who somewhat likely change their support of remote working, with an odds ratio of 0.7222, causing a 27.78% decrease in the odds ratio.

Changes in the barriers to remote work responsibilities, which have significantly worsened for people in their organizations, make them more likely to prefer working in the office compared to those for whom it has somewhat improved, with an odds ratio of 0.5699, causing a 43% decrease in the odds ratio. Similarly, changes in the barriers of the software of the organizations for people for whom it has somewhat worsened make them more likely to work in the office, with an odds ratio of 0.7814, compared to people for whom the software of the organization remained the same.

Some kinds of work cannot be done remotely. People have been asked if work that is not possible to be done remotely is a barrier for them. People who said it is not a barrier for them are more likely to work remotely compared to people for whom it stayed the same, with an odds ratio of 1.5468, causing a 54% increase in the odds ratio for people for whom it is not a barrier. Motivation is also the main factor in remote work or in the office. It has been asked if motivation has improved over the past six months. People who said that it has somewhat improved are more likely to prefer to work in the office compared to people for whom it stayed the same, with an odds ratio of 0.6776, causing a 32% decrease in the odds ratio.

People who strongly agreed that their productivity has somewhat improved are less likely to work remotely compared to people for whom it remained the same, with an odds ratio of 0.7092. However, people who strongly disagreed with remote productivity are more likely to work remotely compared to people for whom it remained the same, with an odds ratio of 1.5135.

Table 2: Table shows the odds ratios of the important variables.
Note for codes e.g. NPNN see the abbreviation table.

	Odds ratios
Intercept	1.8302
IndustryFin & Ins	0.8446
IndustryAdmin-Support	0.6312
IndustryHospitality	0.7658
HouseholdSingle-Family (Other)	0.7349
EmployerOnsiteWorkPolicyNo	1.0357
RemoteWrkEncrgmntStA	1.1950
RemoteCollabEaseStA	1.0046
RemoteCollabEaseSD	0.7501
RemoteCollabEaseStD	0.5226
RemoteImpactEmpSN	0.6712
RemoteImpactEmpSP	0.6226
RemoteImpactEmpNPNN	0.7350
RemoteSuppPartySenior leadership	0.7904
EmpEncrgRemoteVery unlikely	0.5952
EmployerSupportingRemoteChangesVery unlikely	0.7222
ChangeBarriersResposiSigWr	0.5699
OrgSoftwareSW	0.7814
UnremotableTasksNotBerr	1.5468
MotivationSI	0.6776
Productivity_comp_Remote_WplaceSA	0.7092
Productivity_comp_Remote_WplaceStD	1.5135

	Odds ratios
Productivity_comp_Remote_WplaceStA	0.6380

4. Conclusions

In conclusion, our study has shed light on the complex interplay of factors influencing work preferences in the post-COVID-19 world. I found that industry type, household composition, and organizational support significantly impact individuals' choices between remote and office-based work.

Certain industries, like finance and insurance, lean toward traditional office setups, also administrative and support roles exhibit a preference for in-person work. The composition of households also plays a role, with single-family households showing a preference for office work compare to the child-rearing couples.

The support and policies of organizations, changes in barriers to remote work, and individual motivation are critical drivers of these preferences. Recognizing these influences is paramount as organizations adapt to evolving work dynamics. The key point is all those determinants that I found important should be considered. If employer consider both work from or at office, then they should make sure there is no barriers that distract employees. For example, some employers wants their employees to work in the office but employees do not want to work in the office. Employer should find those distractions and determinant that negatively effects employees. In this study, based on the given data, I have found some important determinants that every employer should consider.

This report provides valuable insights to inform decision-making in a fluid work environment. It highlights the importance of tailoring workplace strategies to accommodate the unique needs and preferences of various sectors and individual. In the future, a flexible approach to work arrangements will be key to meeting the diverse demands of the modern workforce.

There are some limitations to this study. I utilized the Metropolis-Hastings algorithm to fit Bayesian logistic regression. I applied the Metropolis-Hastings algorithm to numerous simulations with varying complexities, such as an increasing number of variables. As I increased the number of variables, the tuning parameter of the proposal distribution needed to be decreased in order to achieve the desired acceptance rate. However, this approach has a disadvantage: it requires very small steps to explore the parameter space, which, in turn, necessitates a large number of iterations and can cause high correlation among the draws. In my case, I ran 100,000 iterations, which took 20 hours, but the algorithm still did not fully converged. I had to rely on the results from these 100,000 iterations due to computational limitations. Another limitation is related to the variable selection method, wherein I rely on subjective decisions for selecting the number of variables based on their probabilities of being chosen. I used a threshold of 0.5, but others may opt for different threshold values that can alter the count of important variables.

The first limitation can be resolved by using other Markov chain Monte carlo method the so-called Hamiltonian Monte carlo method that does not use random walk proposal distribution and instead take large steps to reduce correlation among the draws thereby needs less iterations and computational time.

As metropolis hasting was a bit challenging in our case having 196 variables. I have been going through different methods of whether metropolis can be improved to take less computational time. As a result, I found that Hamiltonian Monte carlo method will be perfect in this situation by taking large steps to draw less correlated variables and reach the desired parameter space quickly. Unfortunately, because of the less time, I was not able to implement that Hamiltonian Monte carlo simulation here.

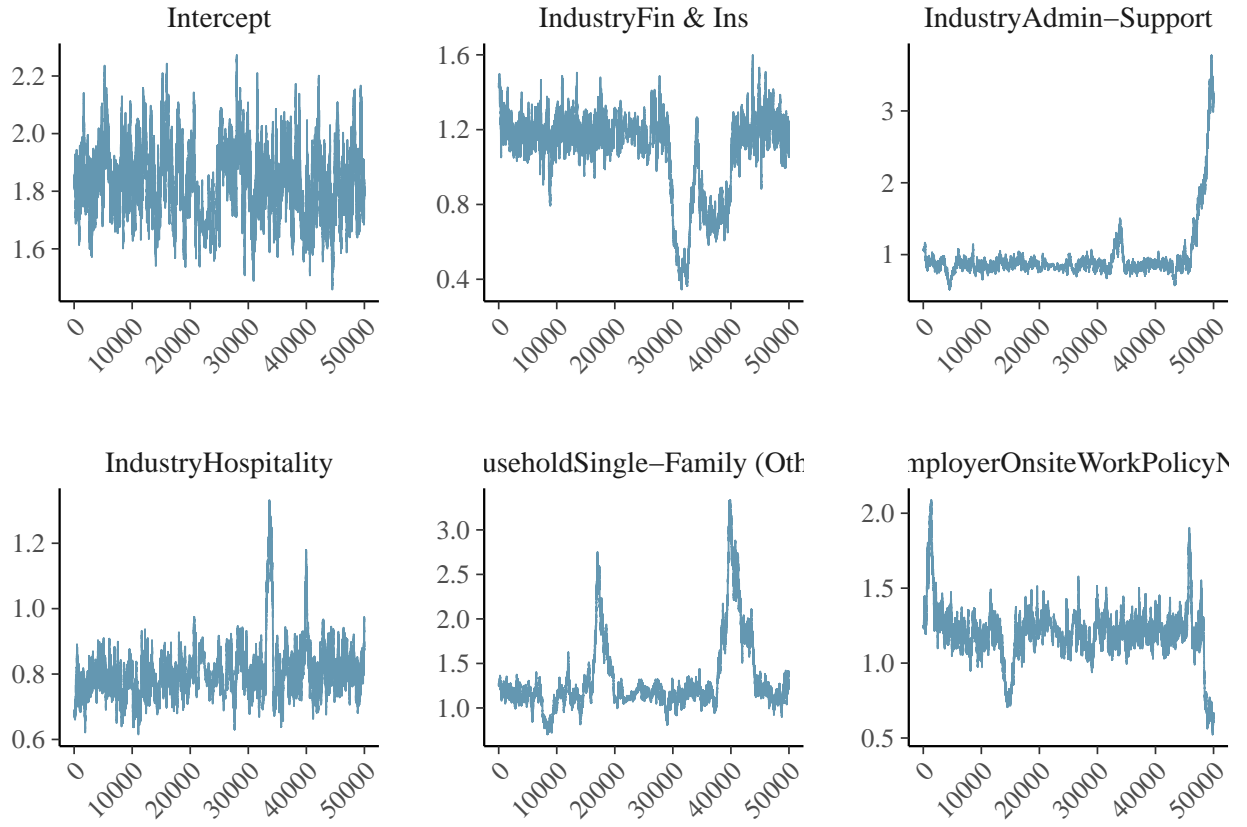


Figure 1: Trace plot of the draws of some of the important variables.

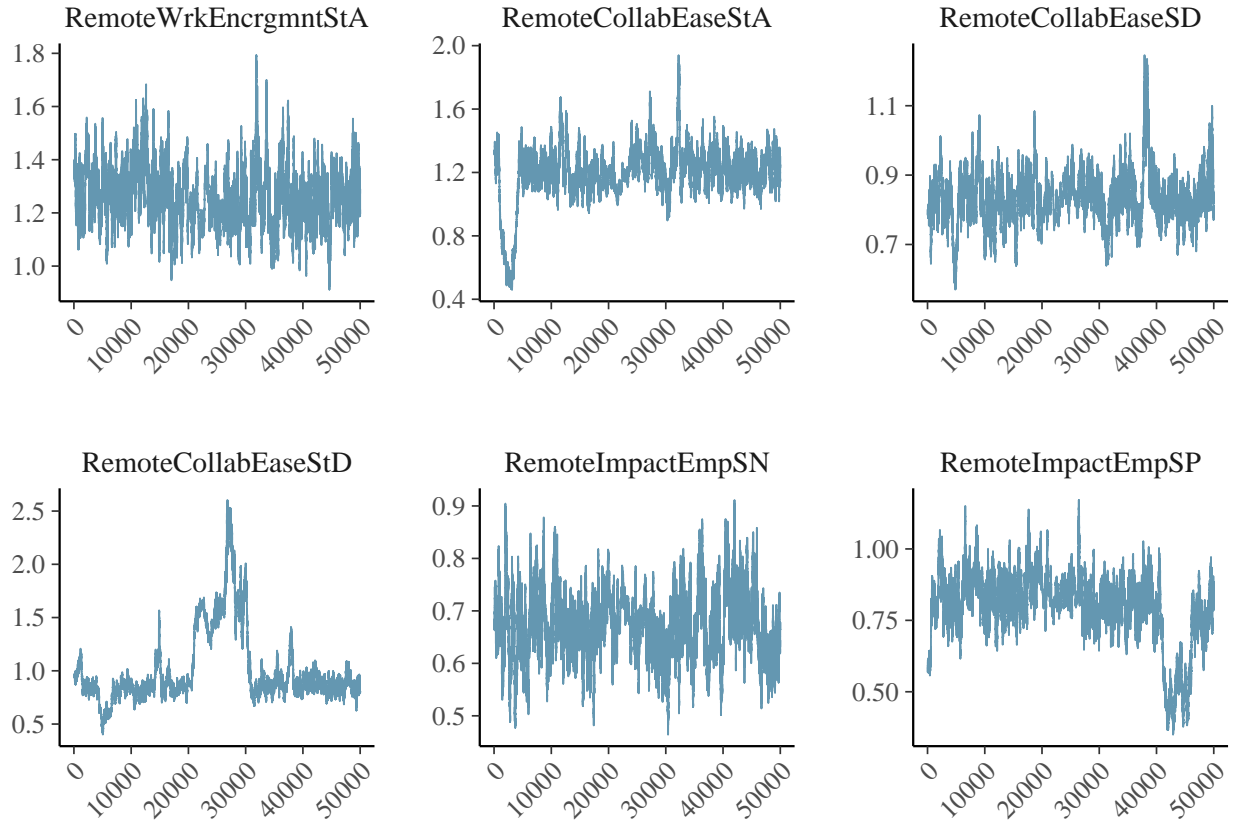


Figure 2: Trace plot of the draws of some of the important variables coefficients.

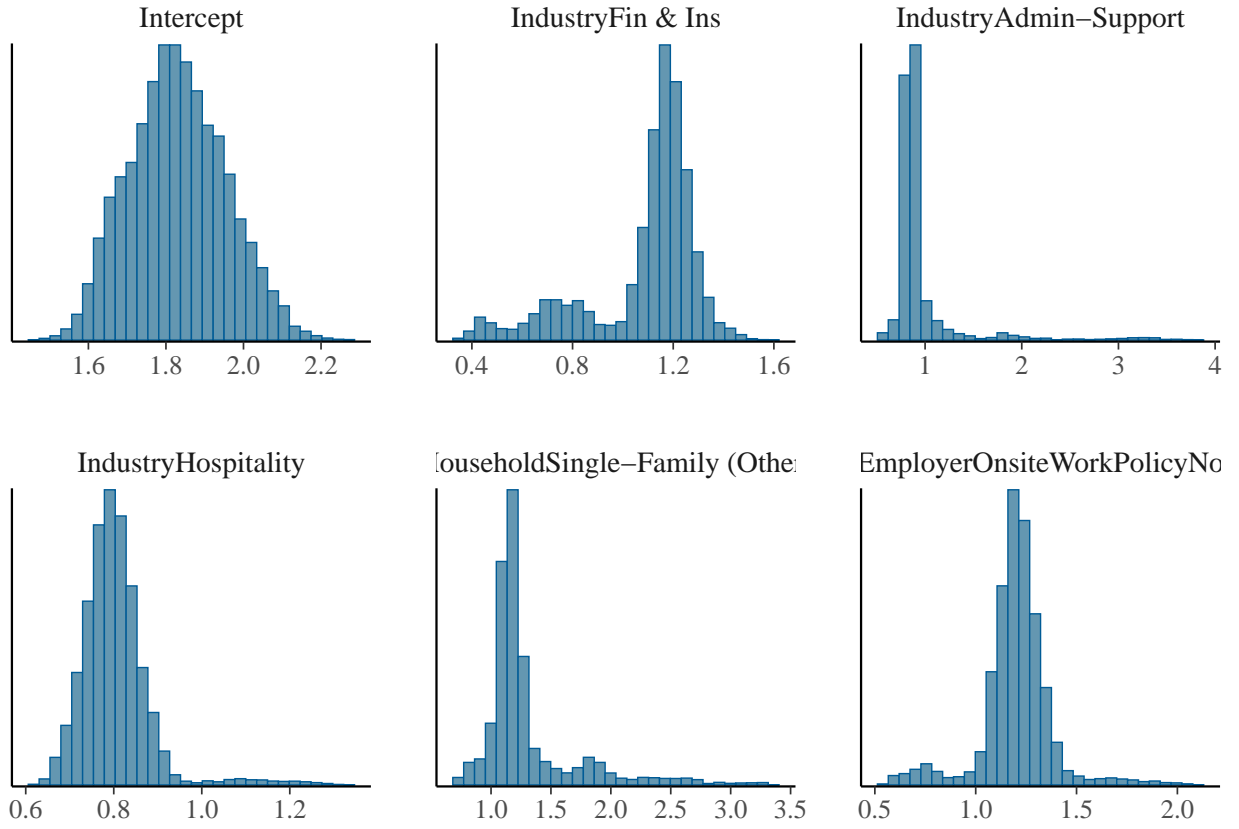


Figure 3: Distribution of the draws of some of the important variables coefficients.

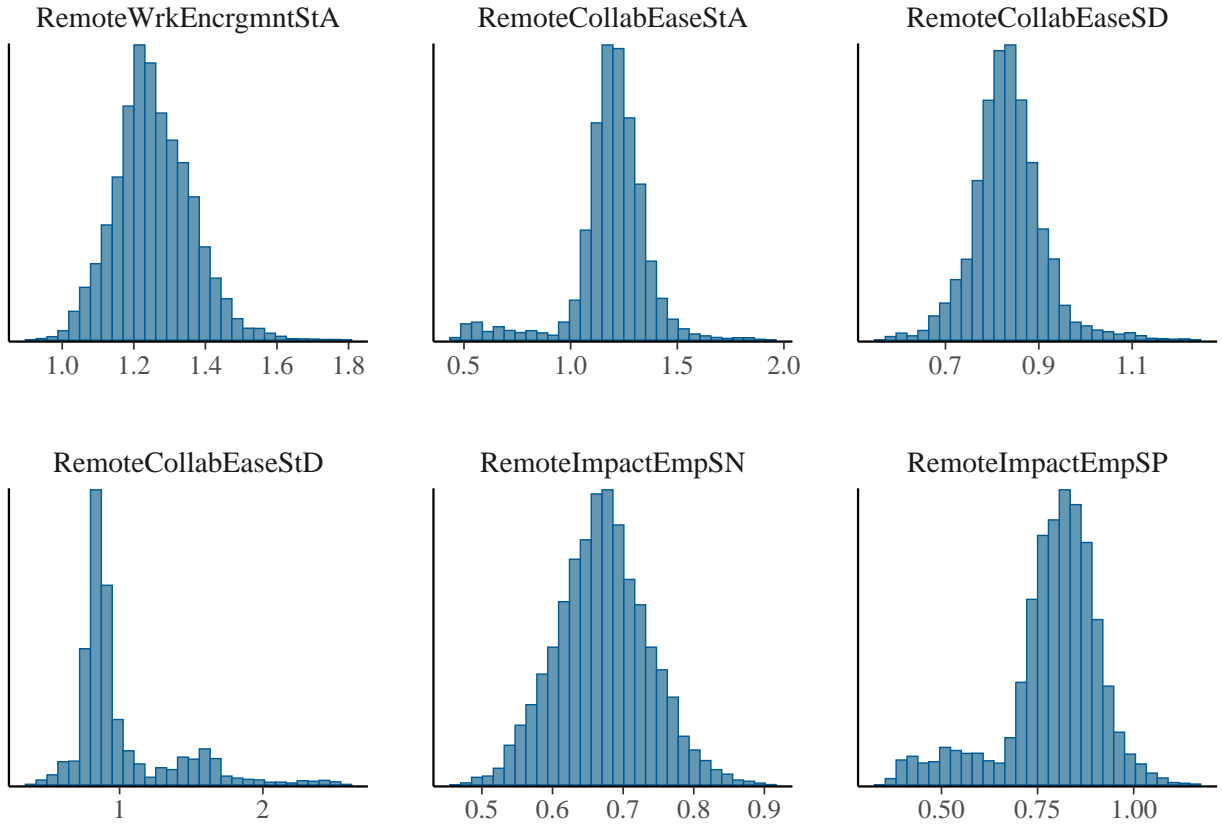


Figure 4: Distribution of the draws of the coefficients of some of the important variables coefficients.

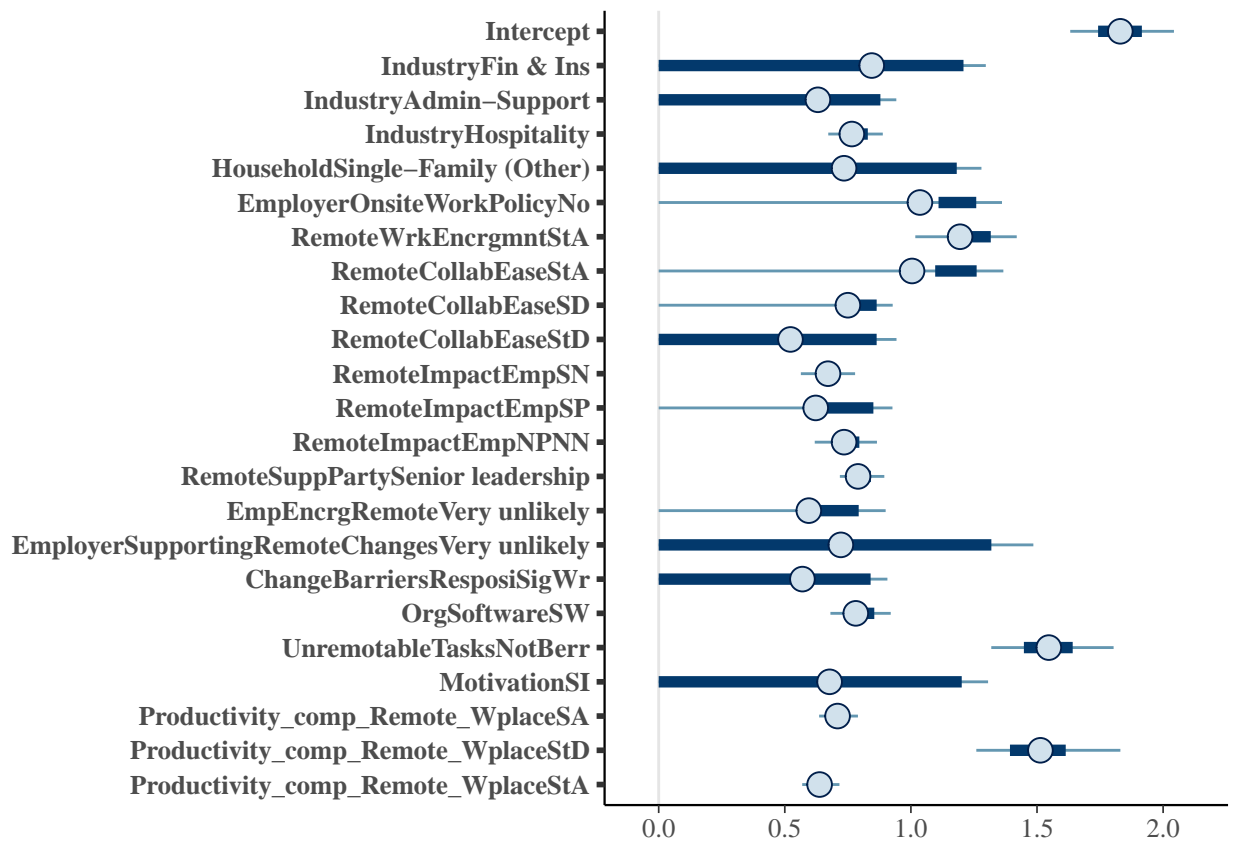


Figure 5: Plot of the Odds ratios of the variables with their respective confidence intervals taking into account when the variable is active or not.

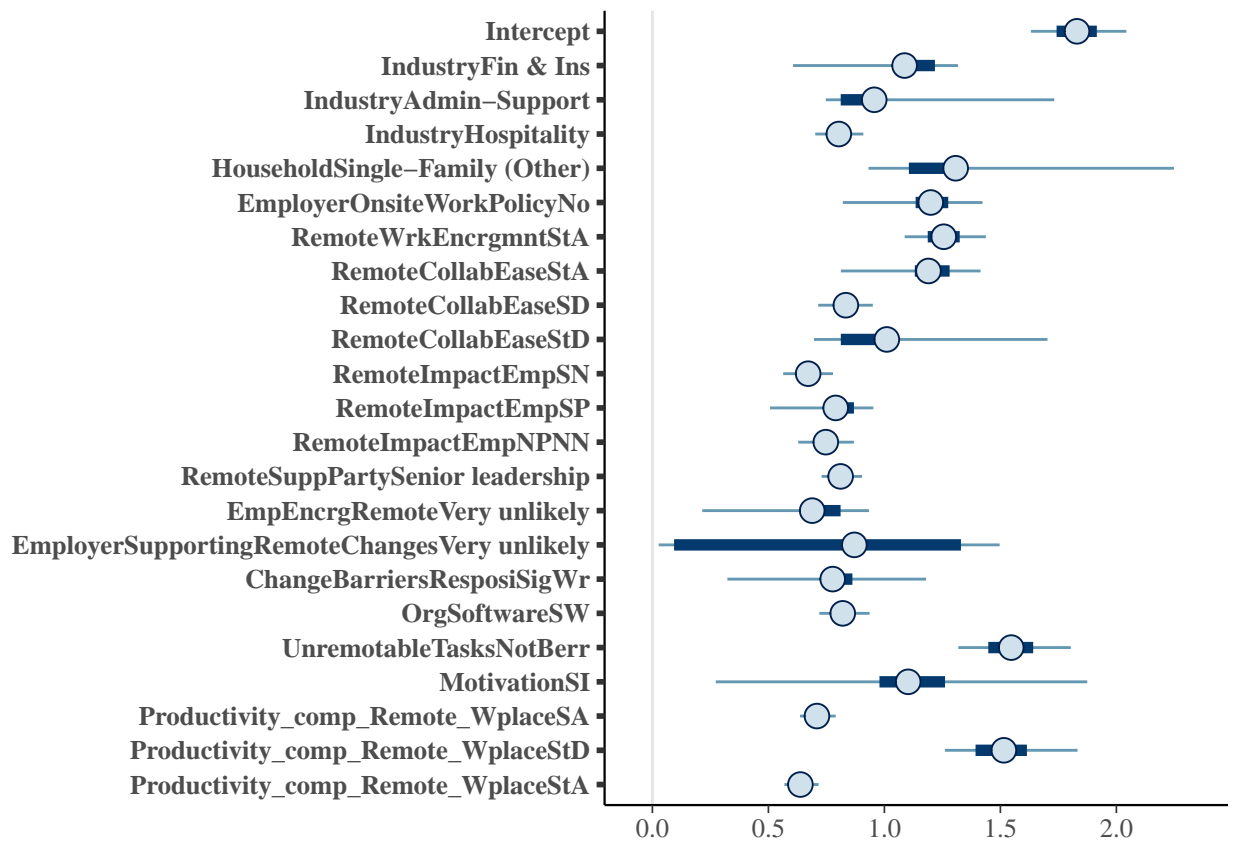


Figure 6: Plot of the odds ratios of the variables with their respective confidence intervals without condering variable selection.

Appendix

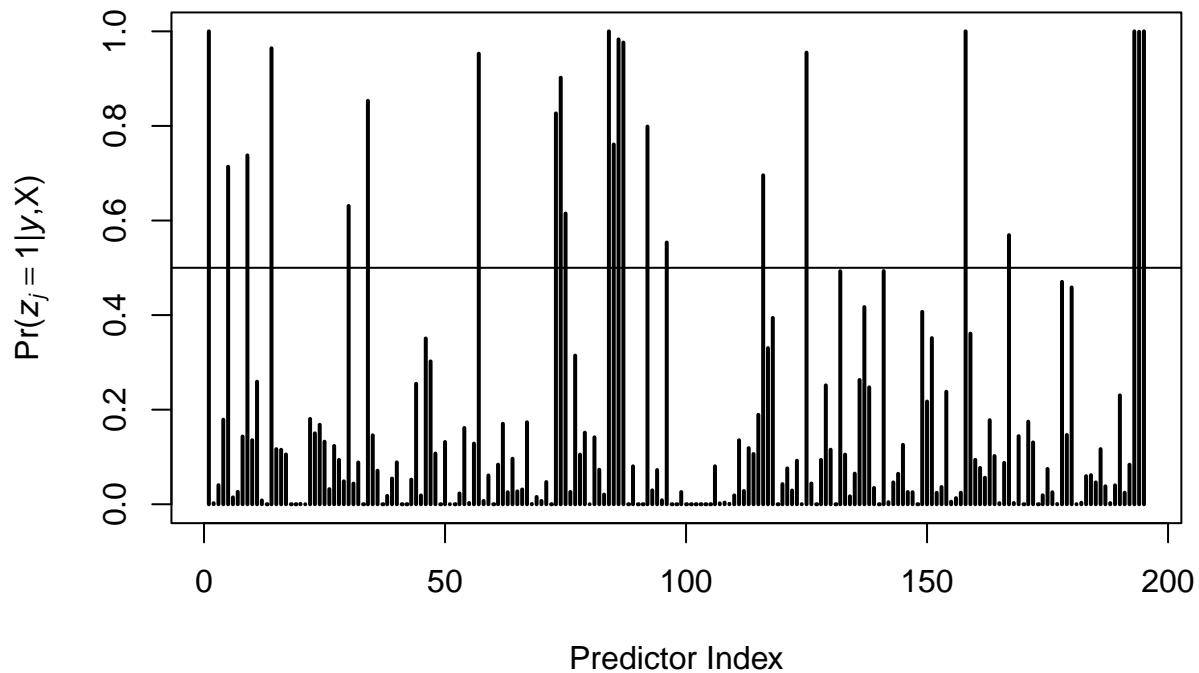


Figure 7: Variable selection.

```
library(tidyverse)
library(MCMCpack)
library(MASS)
library(bayesplot)

#### Loading data

wfh <- read.csv("2021_rws.csv", header = T, check.names = F)

#### Data cleaning

newnames <- c("ResponseID", "BirthYear", "Gender",
              "JobTenure", "Industry", "Occupation",
              "NumEmployees", "Household", "Location",
              "RemoteWorkQ4_2020", "PreferredRemoteQ4_2020",
              "RemoteWorkPercentageYTD_5DayWeek",
              "PreferredRemotePercentageYTD_5DayWeek",
```

"FuturePreferredRemote_5DayWeek",
 "COVIDPolicyChange",
 "PartRemoteWorkLast6Months",
 "FrequencyPartRemoteWork",
 "PreferredFrequencyPartRemoteWork",
 "EmployerOnsiteWorkPolicy",
 "RequiredOnsiteTimePercentage",
 "RemotePolicySuitsMe", "RemoteWorkAutonomy",
 "RemoteWorkDayChoice", "ManagerRemoteWorkDiscretion",
 "OverallRemoteWorkSentiment", "RemoteWrkEncrgmnt",
 "OrganizationRemoteWorkPreparedness", "CommonalityOfRemoteWork",
 "EaseOfRemoteWorkPermission", "RemoteCollabEase",
 "RemotePromoImpact", "RemoteWorkBreakFrequency",
 "RemoteImpactEmp", "RemoteSuppParty",
 "EmpEncrgRemote", "EmployerSupportingRemoteChanges",
 "PersonalChoice", "TimeSpentOnCommuteOnWorkday",
 "TimeSpentWorkingOnWorkday", "TimeSpentOnCaringAndDomesticTasksOnWorkday",
 "WorkplacePersonalFamilyTimeOnWorkday", "WorkplaceSleepOnWorkday",
 "RemoteCommuteTime", "RemoteWorkWorkingTime",
 "RemoteWorkCaringAndDomesticTime", "RemoteWorkPersonalFamilyTime",
 "RemoteWorkSleep", "InterestInNegotiatingPayCutForRemote",
 "MaxAcceptedPayCutPercentage", "ChangeBarriersResposi",
 "ChangeBarr_connect", "OrgSoftware",
 "CollabDifficlty", "PoorMngmnt",
 "ITEquipment", "IsolationFeeling",
 "ExtraCosts", "CyberSecurity",
 "UnremotableTasks", "Workspace",
 "Motivation",
 "Barriers_Remote_Management_Discouragement",
 "Living_Situation", "Lack_Skills",
 "Health_Safety",
 "Barriers_Workplace_Biggest_Noise_Interruptions_Collaboration_Privacy",
 "Barriers_Workplace_Smallest_Noise_Interruptions_Collaboration_Privacy",
 "Barriers_Workplace_Biggest_Connectivity_Workspace_AV_Setup_Privacy",
 "Barriers_Workplace_Smallest_Connectivity_Workspace_AV_Setup_Privacy",
 "Barriers_Workplace_Biggest_Interruptions_Connectivity_Workspace_Innovation",
 "Barriers_Workplace_Smallest_Interruptions_Connectivity_Workspace_Innovation",
 "Barriers_Workplace_Biggest_Connectivity_Collaboration_Innovation_Privacy",
 "Barriers_Workplace_Smallest_Connectivity_Collaboration_Innovation_Privacy",
 "Barriers_Workplace_Biggest_Noise_Interruptions_Connectivity_Innovation",
 "Barriers_Workplace_Smallest_Noise_Interruptions_Connectivity_Innovation",
 "Barriers_Workplace_Biggest_Workspace_AV_Setup_Collaboration_Innovation",
 "Barriers_Workplace_Smallest_Workspace_AV_Setup_Collaboration_Innovation",
 "Barriers_Workplace_Biggest_Interruptions_Workspace_Collaboration_Privacy",
 "Barriers_Workplace_Smallest_Interruptions_Workspace_Collaboration_Privacy",
 "Barriers_Workplace_Biggest_Noise_AV_Setup_Collaboration_Innovation",
 "Barriers_Workplace_Smallest_Noise_AV_Setup_Collaboration_Innovation",
 "Agree_Disagree_Remote_Work_Better_Feel",
 "Agree_Disagree_Remote_Work_Better_Active",
 "Agree_Disagree_In_Person_Colleagues_Feel_Better",
 "Manage_People_Current_Job",
 "Employees_You_Manage_Work_Remotely",
 "Agree_Disagree_Discretion_Offer_Deny_Remote_Work_Employees_Manage",


```

    "Agree_Disagree_Remote_Work_Retain_Employees",
    "Agree_Disagree_Remote_Work_Recruit_Employees",
    "Agree_Disagree_Team_Works_Well_Remotely",
    "Agree_Disagree_Easy_Manage_Remote_Employees",
    "Agree_Disagree_Easy_Manage_Poor_Performers_Remotely",
    "Agree_Disagree_Feel_Well_Prepared_Manage_Remotely",
    "Agree_Disagree_Managing_Remotely_Focused_On_Results",
    "Agree_Disagree_Easy_Contact_Employees_Remotely",
    "Productivity_comp_Remote_Wplace",
    "Productivity_Comparison_Employees_You_Manage_Remote",
    "Employees_Work_Remotely",
    "Discretion_Remote_Work_Management",
    "Remote_Work_Retention",
    "Remote_Work_Recruitment",
    "Team_Collaboration_Remote_Work",
    "Ease_Manage_Remote_Employees",
    "Ease_Manage_Poor_Performers_Remote",
    "Preparedness_Manage_Remote_Employees",
    "Focus_On_Results_Remote_Management",
    "Ease_Contact_Remote_Employees",
    "Productivity_Comparison_Self",
    "Productivity_Comparison_Managed_Employees")

wfh <- wfh %>% rename_at(vars(colnames(.)), ~ newnames) %>%
  #select(3:60) %>%
  mutate(across(where(is.character), as_factor)) %>%
  mutate(weights_resp = fct_recode(FuturePreferredRemote_5DayWeek,
    ".8" = "80%",
    ".5" = "50% - About half of my time",
    ".6" = "60%",
    ".4" = "40%",
    "1" = "100% - All of my time",
    ".9" = "90%",
    ".5" = "Less than 10% of my time",
    "0" = "I would prefer not to work remotely",
    ".2" = "20%",
    ".1" = "10%",
    ".3" = "30%",
    ".7" = "70%")) %>%
  mutate(weights_resp = as.character(weights_resp)) %>%
  mutate(weights_resp = as.numeric(weights_resp)) %>%
  mutate(JobTenure = fct_recode(JobTenure, "2" = "Between 1 and 5 years", #1-5 Yrs
    "1" = "Between 6 and 12 months", #6-12 Mths
    "3" = "More than 5 years"), #5+ Yrs
    JobTenure = as.numeric(JobTenure),
    weights_resp = as.numeric(weights_resp),
    FuturePreferredRemote_5DayWeek = fct_recode(FuturePreferredRemote_5DayWeek,
      "Yes" = "80%",
      "Yes" = "50% - About half of my time",
      "Yes" = "60%",

```

```

        "No" = "40%",
        "Yes" = "100% - All of my time",
        "Yes" = "90%",
        "No" = "Less than 10% of my time",
        "No" = "I would prefer not to work remotely",
        "No" = "20%",
        "No" = "10%",
        "No" = "30%",
        "Yes" = "70%"),
Industry = fct_recode(Industry, "Pro, Sci, Tech" = "Professional, Scientific and Technical",
        "Fin & Ins" = "Financial and Insurance",
        "Const" = "Construction",
        "Health-Social Care" = "Health Care and Social Assistance",
        "Manuf" = "Manufacturing",
        "Pub Admin & Safety" = "Public Administration and Safety",
        "Admin-Support" = "Administrative and Support",
        "Utility Sector" = "Electricity, Gas, Water and Waste",
        "E&T Sector" = "Education and Training",
        "Retail" = "Retail Trade",
        "Hospitality" = "Accommodation and Food",
        "Wholesale" = "Wholesale Trade",
        "Telecom" = "Information Media and Telecommunications",
        "Property Management" = "Rental, Hiring and Real Estate",
        "Transport & Logistics" = "Transport, Postal and Warehousing",
        "Agri, Forestry & Fishing" = "Agriculture, Forestry and Fishing",
        "Arts & Recreation" = "Arts and Recreation",
        "Mining" = "Mining"),
Household = fct_recode(Household, "Child-Rearing Couple" = "Couple with dependent children",
        "Couple Alone" = "Couple with no dependent children",
        "Single" = "Single person",
        "Solo Parent Family" = "One parent family with dependent children",
        "Co-living" = "Group household",
        "Single-Family (Other)" = "Other one family household",
        "Multi-Household" = "Multiple family household"),
#FrequencyPartRemoteWork = fct_recode(FrequencyPartRemoteWork, "0%" = ""),
#PreferredFrequencyPartRemoteWork = fct_recode(PreferredFrequencyPartRemoteWork, "0%" = "")
RemotePolicySuitsMe = fct_recode(RemotePolicySuitsMe, "SA" = "Somewhat agree",
        "NAND" = "Neither agree nor disagree",
        "StA" = "Strongly agree",
        "StD" = "Strongly disagree",
        "SD" = "Somewhat disagree"),
RemoteWorkDayChoice = fct_recode(RemoteWorkDayChoice, "SA" = "Somewhat agree",
        "NAND" = "Neither agree nor disagree",
        "StA" = "Strongly agree",
        "StD" = "Strongly disagree",
        "SD" = "Somewhat disagree"),
RemoteWrkEncrgmnt = fct_recode(RemoteWrkEncrgmnt, "SA" = "Somewhat agree",
        "NAND" = "Neither agree nor disagree",
        "StA" = "Strongly agree",
        "StD" = "Strongly disagree",
        "SD" = "Somewhat disagree"),
RemoteCollabEase = fct_recode(RemoteCollabEase, "SA" = "Somewhat agree",
        "NAND" = "Neither agree nor disagree",

```

```

        "StA" = "Strongly agree",
        "StD" = "Strongly disagree",
        "SD" = "Somewhat disagree"),
RemotePromoImpact = fct_recode(RemotePromoImpact, "SA" = "Somewhat agree",
        "NAND" = "Neither agree nor disagree",
        "StA" = "Strongly agree",
        "StD" = "Strongly disagree",
        "SD" = "Somewhat disagree"),
RemoteImpactEmp = fct_recode(RemoteImpactEmp, "StP" = "Strongly Positive",
        "SN" = "Somewhat Negative",
        "SP" = "Somewhat Positive",
        "NPNN" = "Neither positive nor negative",
        "SN" = "Strongly Negative"),
PersonalChoice = fct_recode(PersonalChoice, "SL" = "Somewhat likely",
        "NUOL" = "Neither unlikely or likely",
        "VL" = "Very likely",
        "SU" = "Somewhat unlikely",
        "VU" = "Very unlikely"),
ChangeBarriersResposi = fct_recode(ChangeBarriersResposi, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
ChangeBarr_connect = fct_recode(ChangeBarr_connect, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
OrgSoftware = fct_recode(OrgSoftware, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
CollabDifficlty = fct_recode(CollabDifficlty, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
PoorMngmnt = fct_recode(PoorMngmnt, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
ITEquipment = fct_recode(ITEquipment, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",

```

```

        "SigWr" = "Significantly worsened"),
IsolationFeeling = fct_recode(IsolationFeeling, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
ExtraCosts = fct_recode(ExtraCosts, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
CyberSecurity = fct_recode(CyberSecurity, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
UnremovableTasks = fct_recode(UnremovableTasks, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
Workspace = fct_recode(Workspace, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
Motivation = fct_recode(Motivation, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
Living_Situation = fct_recode(Living_Situation, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
Lack_Skills = fct_recode(Lack_Skills, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",
        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
Health_Safety = fct_recode(Health_Safety, "SI" = "Somewhat improved",
        "StaySame" = "Stayed about the same",
        "NotBerr" = "Not a barrier for me",
        "SW" = "Somewhat worsened",

```

```

        "SigIm" = "Significantly improved",
        "SigWr" = "Significantly worsened"),
  Productivity_comp_Remote_Wplace = fct_recode(Productivity_comp_Remote_Wplace,
        "SA" = "Somewhat agree",
        "NAND" = "Neither agree nor disagree",
        "StA" = "Strongly agree",
        "StD" = "Strongly disagree",
        "SD" = "Somewhat disagree")
) %>%
mutate_at(.vars = vars(FrequencyPartRemoteWork,
  PreferredFrequencyPartRemoteWork,
  RequiredOnsiteTimePercentage),
  .funs = forcats::fct_recode,
  "Zero%" = "") %>%
dplyr::select(-c(RemoteWorkQ4_2020, PreferredRemoteQ4_2020,
  Occupation, RemoteWorkPercentageYTD_5DayWeek,
  PreferredRemotePercentageYTD_5DayWeek,
  MaxAcceptedPayCutPercentage, BirthYear,
  ResponseID,
  PartRemoteWorkLast6Months,
  FrequencyPartRemoteWork,
  PreferredFrequencyPartRemoteWork,
  Productivity_Comparison_Employees_You_Manage_Remote,
  Employees_Work_Remotely, Discretion_Remote_Work_Management,
  Remote_Work_Retention, Remote_Work_Recruitment,
  Productivity_Comparison_Managed_Employees,
  Productivity_Comparison_Self,
  Team_Collaboration_Remote_Work, Ease_Manage_Remote_Employees,
  Ease_Manage_Poor_Performers_Remote,
  Preparedness_Manage_Remote_Employees, Focus_On_Results_Remote_Management,
  Ease_Contact_Remote_Employees,
  Agree_Disagree_Managing_Remotely_Focused_On_Results,
  Agree_Disagree_Easy_Contact_Employees_Remotely)) %>%
dplyr::select(-where(~sum(is.na(.)) >= 50)) %>%
dplyr::select(-c("RequiredOnsiteTimePercentage")) %>%
dplyr::select(-c(weights_resp))

#####
#### Metropolis Hastings
#####

df <- wfh %>%
  na.omit() # removed 101 na values

X <- as.data.frame(model.matrix(~., data = df[, -c(7)])) %>% # design matrix
  dplyr::select(-c("(Intercept)"))

X <- as.data.frame(cbind(1, scale(X))) # scaling design matrix

```

```

p <- dim(X)[2] # number of variables
n <- dim(X)[1] # number of samples

Y <- rep(0,n)
Y[df$FuturePreferredRemote_5DayWeek == "Yes"] <- 1

iterations <- 100000

# for prior distribution
b0 <- matrix(rep(0, p), nrow = p) #apply(X, 2, mean)
SIGMA <- diag(100,nrow = p)

## Response Y is binary
## I will be using logistic regression

### Loglikelihood
loglikli <- function(betas, X){
  sum((as.matrix(X)%*%betas) * Y -
      apply(as.matrix(X)%*%betas, 1, function(x) log(1 + exp(x))))
}

## Prior
proir_dis <- function(betas, b0, SIGMA){
  (-1/2)*crossprod((betas - b0), solve(SIGMA))%*%(betas - b0)
}

## Posterior
target_pdf <- function(betas_current, betas_prime, b0, SIGMA, X){
  target_dis <- exp(loglikli(betas_prime, X) +
                    proir_dis(betas_prime, b0, SIGMA)[1,1] -
                    loglikli(betas_current, X) -
                    proir_dis(betas_current, b0, SIGMA)[1,1])
  return(target_dis)
}

z <- c(1, rbinom(p-1, 1, .5))

MH <- function(iterations, b0, SIGMA) {
  BETAs <- Z <- matrix(NA, nrow = iterations, ncol = p)
  Accept <- rep(0, iterations)
  fit <- glm(Y~-1+., family = binomial(link = "logit"),
            data = X)
  betas_current <- as.matrix(fit$coefficients, nrow= p)
  glm_cov <- summary(fit)$cov.unscaled # made changes

```

```

for (i in 1:iterations) {
  lpy.c<-loglikli(betas_current[z==1], X[,z==1,drop=FALSE])
  for (j in sample(2:p)) {
    zp <- z
    zp[j] <- 1 - zp[j] # Switch value of z_j
    lpy.p<- loglikli(betas_current[zp==1], X[,zp==1,drop=FALSE])

    r <- (lpy.p - lpy.c) * (-1)^(zp[j] == 0)
    z[j] <- rbinom(1, 1, 1 / (1 + exp(-r)))

    if (z[j] == zp[j]) {
      lpy.c <- lpy.p
    }
  }
  #beta <- z
  betas_prime <- as.matrix(mvrnorm(n = 1,
                                   betas_current,
                                   .02*solve((solve(SIGMA) + # beta removed
                                                solve(glm_cov))))), nrow = p)

  acceptance_ratio <- min(1, target_pdf(betas_current[z==1],
                                         betas_prime[z==1], b0[z==1],
                                         SIGMA[z==1,z==1],X[, z == 1, drop = FALSE]))

  u <- runif(1)

  if (u < acceptance_ratio) {
    betas_current <- betas_prime
    Accept[i] <- 1
  }
  BETAs[i, ] <- betas_current
  Z[i, ] <- z
}

return(list(Bs = BETAs, Zs = Z, AcceptRate = mean(Accept)))
}

MH_results <- MH(iterations=iterations, b0, SIGMA)

# As I'm using rmarkdown. I saved results of the metropolis
# Hastings so that I can rerun rmarkdown file again and again
# without taking much time in running metropolis Hastings

load("C:/Users/MyLaptop/OneDrive/Atiq/Bayesian/Project/MH_results.RData")

burnin <- iterations/2
colnames(X)[1] <- "Intercept"
colnames(MH_results$Zs) <- colnames(X)
important_variables <- data.frame(Imp = apply(MH_results$Zs[c(burnin+1):iterations,],2,

```

```

mean)[apply(MH_results$Zs[c(burnin+1):iterations,],2, mean)>.5]]

colnames(MH_results$Bs) <- colnames(X)

post_draws <- as.data.frame(exp(MH_results$Bs))
prob <- as.data.frame(MH_results$Zs)

Imp_var_draws <- post_draws %>%
  dplyr::select(all_of(rownames(important_variables)))

prob <- prob %>%
  dplyr::select(all_of(rownames(important_variables)))

mcmc_trace(Imp_var_draws[c(burnin+1):iterations,1:6],
  pars = c(colnames(Imp_var_draws)[1],
    colnames(Imp_var_draws)[2],
    colnames(Imp_var_draws)[3],
    colnames(Imp_var_draws)[4],
    colnames(Imp_var_draws)[5],
    colnames(Imp_var_draws)[6])) +
  theme(axis.text.x=element_text(angle=45,hjust=1))

mcmc_trace(Imp_var_draws[c(burnin+1):iterations,7:12],
  pars = c(colnames(Imp_var_draws)[7],
    colnames(Imp_var_draws)[8],
    colnames(Imp_var_draws)[9],
    colnames(Imp_var_draws)[10],
    colnames(Imp_var_draws)[11],
    colnames(Imp_var_draws)[12])) +
  theme(axis.text.x=element_text(angle=45,hjust=1))
mcmc_hist(Imp_var_draws[c(burnin+1):iterations,1:6],
  pars = c(colnames(Imp_var_draws)[1],
    colnames(Imp_var_draws)[2],
    colnames(Imp_var_draws)[3],
    colnames(Imp_var_draws)[4],
    colnames(Imp_var_draws)[5],
    colnames(Imp_var_draws)[6]))

mcmc_hist(Imp_var_draws[c(burnin+1):iterations,7:12],
  pars = c(colnames(Imp_var_draws)[7],
    colnames(Imp_var_draws)[8],
    colnames(Imp_var_draws)[9],
    colnames(Imp_var_draws)[10],
    colnames(Imp_var_draws)[11],
    colnames(Imp_var_draws)[12]))

# curve(fitted_Y,
#       from = min(fitted_Y), to = max(fitted_Y),

```



```

#      col = "red", lwd = 2, add = TRUE)
odds_ratios_draws <- as.matrix(prob[c(burnin+1):iterations,
                                  ])*as.matrix(imp_var_draws[c(burnin+1):iterations,])

odds_ratios <- as.data.frame(apply(odds_ratios_draws, 2, mean))

knitr::kable(odds_ratios,
             digits = 4,
             col.names = "Odds ratios",
             caption = "Table shows the odds ratios of the important variables. Note for codes e.g. NPN")

mcmc_intervals(odds_ratios_draws,
               pars = c(colnames(imp_var_draws)), point_est = "mean")

mcmc_intervals(imp_var_draws[c(burnin+1):iterations,],
               pars = c(colnames(imp_var_draws)), point_est = "mean")
plot(apply(MH_results$Zs[c(burnin+1):iterations,], 2, mean), type = "h",
     ylab=expression(paste( "Pr(", italic(z[j] == 1), "|", italic(y), ",X)", sep="")),
     xlab = "Predictor Index",
     lwd=2)
abline(h = 0.5)

```