How are Coloradans Coping with Covid-19?

An Examination of The Factors Associated with Feelings of Depressing During the Coronavirus Pandemic

By: Nick Koprowicz
December 4, 2020

Introduction

Coronavirus Disease 2019 (COVID-19) is a respiratory disease that can be deadly for many whom it infects. As of November 2020, there have been 11.6 million cases and more than 250,000 deaths from COVID-19 in the U.S. In order to combat the spread of the virus, cities and states implemented a variety of lockdown measures, starting in March, during which time only essential businesses were allowed to remain open. Many schools and workplaces switched from in-person to remote operation to minimize person-to-person contact. Some of these restrictions were relaxed during the Summer of 2020 but were ramped up in the Fall as infection rates increased. As a consequence, negative impacts of the pandemic occur on at least three fronts: the physical impacts of the virus itself, the stress and anxiety caused by worry about potential health effects of the virus, and the impact of lockdown measures such as economic hardship and social isolation. These stressors can have a negative impact on individuals' mental health.

In this project, we examine what factors may be contributing to more frequent feelings of depression among individuals in Colorado during the COVID-19 pandemic. We hope that by identifying these factors, we can recommend groups toward which we should direct mental health resources throughout the duration of the pandemic.

Related Work

Though COVID-19 has only impacted the U.S. for several months, researchers have started to study the effects of the pandemic on mental health by using self-reported surveys. In a survey of 3052 adults in April 2020, Jacob Meyer et al. found that lower levels of physical activity, increased screen time, and self-isolation/quarantine during the pandemic were associated with worse depression, loneliness, and stress (Meyer, 2020). Yanmengqian Zhou et al. used online surveys in a longitudinal study between April and May to identify a modest negative impact of the pandemic on mental health, especially among younger adults and those with preexisting conditions (Zhou, 2020). Calliope Holingue et al. analyzed a survey from the Pew Research Center and found that higher reported distress levels were associated with major changes in personal life, perceived economic threat, individual health, and financial uncertainty (Holingue, 2020). Many individuals suffering from mental distress also report difficulty eating and sleeping, and in some cases, increased substance abuse (Panchal, 2020).

In this project, we aim to more fully understand what factors at the individual level such as race and income are impacting mental health during the pandemic. It has been previously shown that during disasters, disadvantaged groups such as racial or ethnic minorities and those with low income experience more severe mental health impacts than their socially advantaged counterparts (Purtle, 2012). By considering these factors in our analysis, we hope to shed some light on the risk level experienced by these groups.

Data

Throughout the pandemic, the U.S. Census Bureau has collected data on household experiences via weekly surveys in order to inform federal and state recovery planning. The *Household Pulse Survey* asks about factors such as employment status, healthcare, housing, and mental health during the COVID-19 pandemic. There have been three phases of the survey, each approximately two months in length, starting in April and set to continue through December. For this project, we use the 16th week of the survey, collected during September 30 – October 12.

In order to understand the factors that impact negative health impacts in Colorado, we consider how respondents answer the prompt: “Frequency of feeling depressed over the previous 7 days”, for which they can select “Not at all”, “Several days”, “More than half the days”, or “Nearly every day”. We redefine this to be a binary response and consider whether respondents reported feeling depressed “not at all” versus “several days or more”. The other factors considered as predictors in our analysis are seen in table 1. Note that many of these factors have been redefined from how they were stated in the *Household Pulse Survey*. For a description of how these factors were originally measured and how we redefined them for our analysis, see table 7 in the appendix. Additionally, we omit from the data respondents who selected “Question seen but category not selected” or “Missing / Did not report” for any question. We do not see a meaningful way to impute this data, nor a way to choose a “default” value that would have a minimal impact on our analysis. This preprocessing left us with 1992 responses, and we feel that is a large enough sample from which to proceed.

We make a few observations from exploring the data, seen in figure 1. Individuals who have recently lost work or who have had delayed medical care tend to report the highest rates of feeling depressed for several days or more during the last 7 days. We also see that younger people and those with low income tend to report more frequent feelings of depression. Finally, we note that the data are mostly balanced with regard to the response (1034 out of the 1992 respondents reported feeling depressed several days or more), as well as gender (1130 of the 1992 respondents were female). However, the data are highly unbalanced when it comes to race, where only 343 of the 1992 respondents were non-white.

Methods

We take a Bayesian approach and choose to model the response using logistic regression. We assume that the responses are independent, that the probability that respondent i feels depressed several days or more during the last 7 days is p_i, and model $\log\left(\frac{p_i}{1-p_i}\right)$ as a linear combination of the other factors. We begin model construction by selecting which factors are important for predicting $\log\left(\frac{p_i}{1-p_i}\right)$.

Factor	Description
AGE	The age of the respondent, an integer
GENDER	Indicator variable for whether the respondent is female
RACE	Indicator variable for whether the respondent is non-white
EDUC	Indicator variable for whether the respondent has a college degree
MS	Indicator variable for whether the respondent is married
NUM_KID	The number of children in the respondent's household, an integer
WRKLOSS	Indicator variable for whether the respondent has experienced a recent household job loss
DELAY	Indicator variable for whether the respondent has experienced delayed medical care in last 4 weeks due to the pandemic
RENT	Indicator variable for whether the respondent does not pay rent
INCOME_1 (reference)	Indicator variable for total household income less than $50,000 before taxes
INCOME_2	Indicator variable for total household income between $50,000 and $100,000 before taxes
INCOME_3	Indicator variable for total household income greater than $100,000 before taxes
DOWN (response)	Indicator variable for whether the respondent felt depressed several days or more during the last 7 days

Table 1. Factors considered in our analysis, with descriptions. Note that INCOME_1 is used as a reference category and does not appear explicitly in our model.

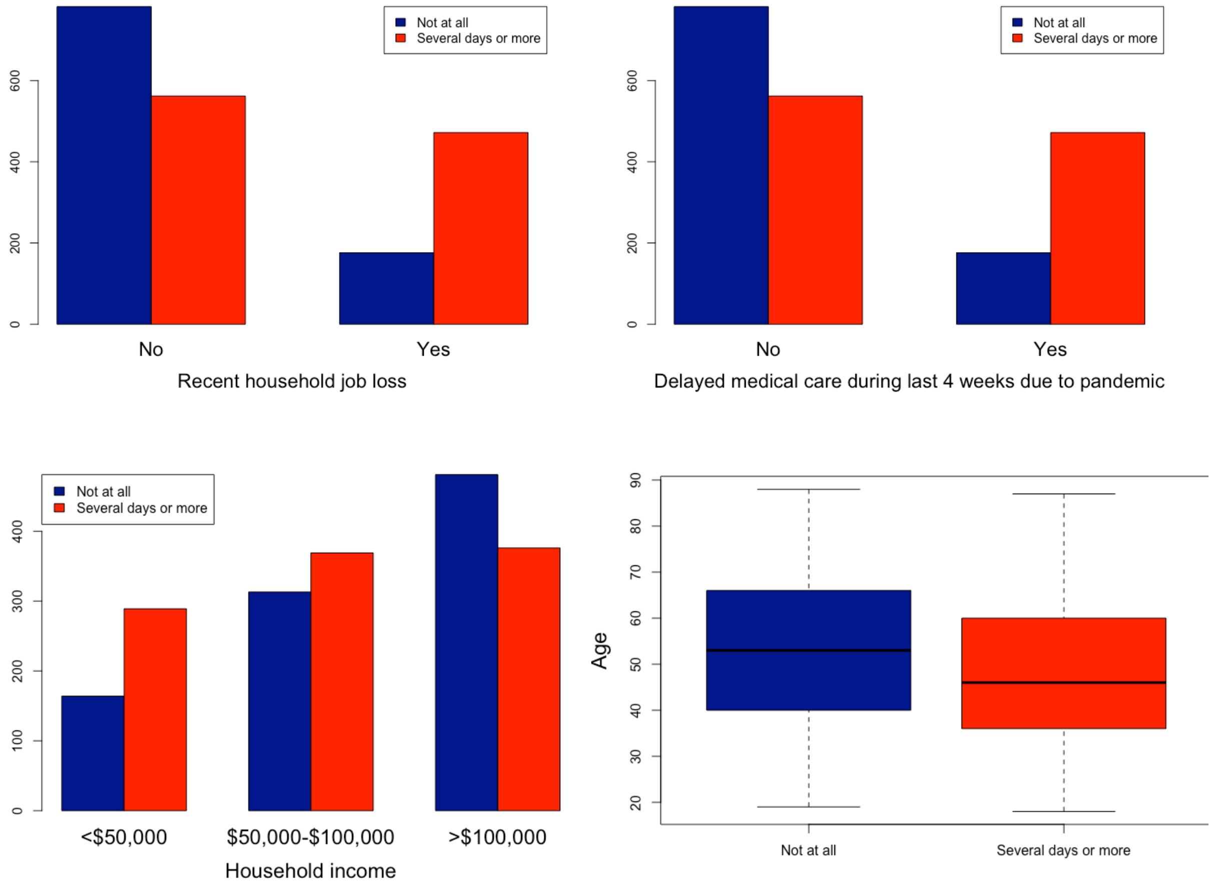

Figure 1. Comparison of the number of respondents who said they felt depressed "several days or more" with other factors. Respondents with recent household job loss, delayed medical care during the last 4 weeks, lower income, and older age report more frequent feelings of depression.

Variable Selection

Due to the large number of factors under consideration, we start the process of variable selection using backward selection with frequentist modeling. Using the `glm` function in R, we fit a logistic regression model using all factors and examine the *p*-values from the summary output. The *p*-values test the null hypothesis that the corresponding factor is not important for predicting the response. We iteratively remove factors with the largest *p*-value, given that the *p*-value is greater than 0.05, until all *p*-values are less than 0.05. Through this process, we eliminate the variables `RACE`, `RENT`, and `EDUC`.

We then fit several Bayes models using the `stanglm` function from the `rstanarm` package in R and compare them using the leave-one-out information criterion (LOOIC) and widely applicable information criterion (WAIC) – two criteria which are based on the estimated log predictive density (ELPD) for new datasets. Starting with the all factors, minus `RACE`, `RENT`, and `EDUC`, we consider candidate models by removing variables that did not appear to have a strong effect on `DOWN` while exploring the data. In each case, we use a skeptical *Normal*(0, 1000) prior for all parameters. We use 4 chains with 10,000 iterations, 5,000 of which are warm-up. Fitting a greater number of iterations in each chain would have been ideal, but because of the large size of the data, each model took approximately 1 minute to fit and would have been too time consuming to use the desired 100,000 iteration length for each candidate model. We check convergence of the chains for each model using the potential scale reduction factors for all parameters, the effective sample size, and Monte Carlo standard error. Our results for the WAIC and LOOIC for each model are summarized in table 2.

Factors removed	WAIC	LOOIC
None (full model minus RACE, RENT, and EDUC)	2470.353	2470.358
GENDER	2471.368	2471.373
MS	2472.465	2472.470
NUM_KIDS	2476.276	2476.280
Age	2503.804	2503.809
GENDER and MS	2473.892	2473.896
GENDER and NUM_KIDS	2477.356	2477.360
MS and NUM_KIDS	2482.964	2482.967
GENDER and MS and NUM_KIDS	2482.964	2482.967

Table 2. Comparison of Bayesian logistic regression models using WAIC and LOOIC. The left column shows factors removed from the full model containing all factors listed in table 1 minus RACE, RENT, and EDUC. We prefer models with smaller WAIC and LOOIC.

Removing GENDER and MS individually, and jointly, have a minimal effect on the WAIC and LOOIC. We also compute the difference in ELPD between the model without GENDER and MS and the full model to be 1.8 with a standard error of 2.8, meaning its plausible that the ELPD between the two models is the same. We decide to remove GENDER and MS from the model, trading a small expense in predictive accuracy for a more parsimonious model.

Final Model

Having selected the important factors, we fit a model with the chosen variables using the base stan package with 4 chains and 100,000 iterations (50,000 of which are warmup) for better posterior estimates. We continue with *Normal*(0, 1000) priors for all parameters. Priors centered at 0 coincide with our assumption that the factors have no effect on the response, and the large variance minimizes our influence in posterior estimates and allows the data to speak for themselves. To check chain convergence, we use the Gelman-Rubin statistics for all variables and see that they are close to 1, we examine trace plots for each variable and see no discernable patterns, and we look at autocorrelation plots and see that autocorrelation delays quickly – all indicating lack of evidence that chains did not converge. These are show in table 3 and figures 2 and 3, respectively. We report the mean, median, 0.025 quantile and 0.975 quantile for all coefficients in table 4. Using the posterior means, a pointwise estimate of our model is in figure 4.

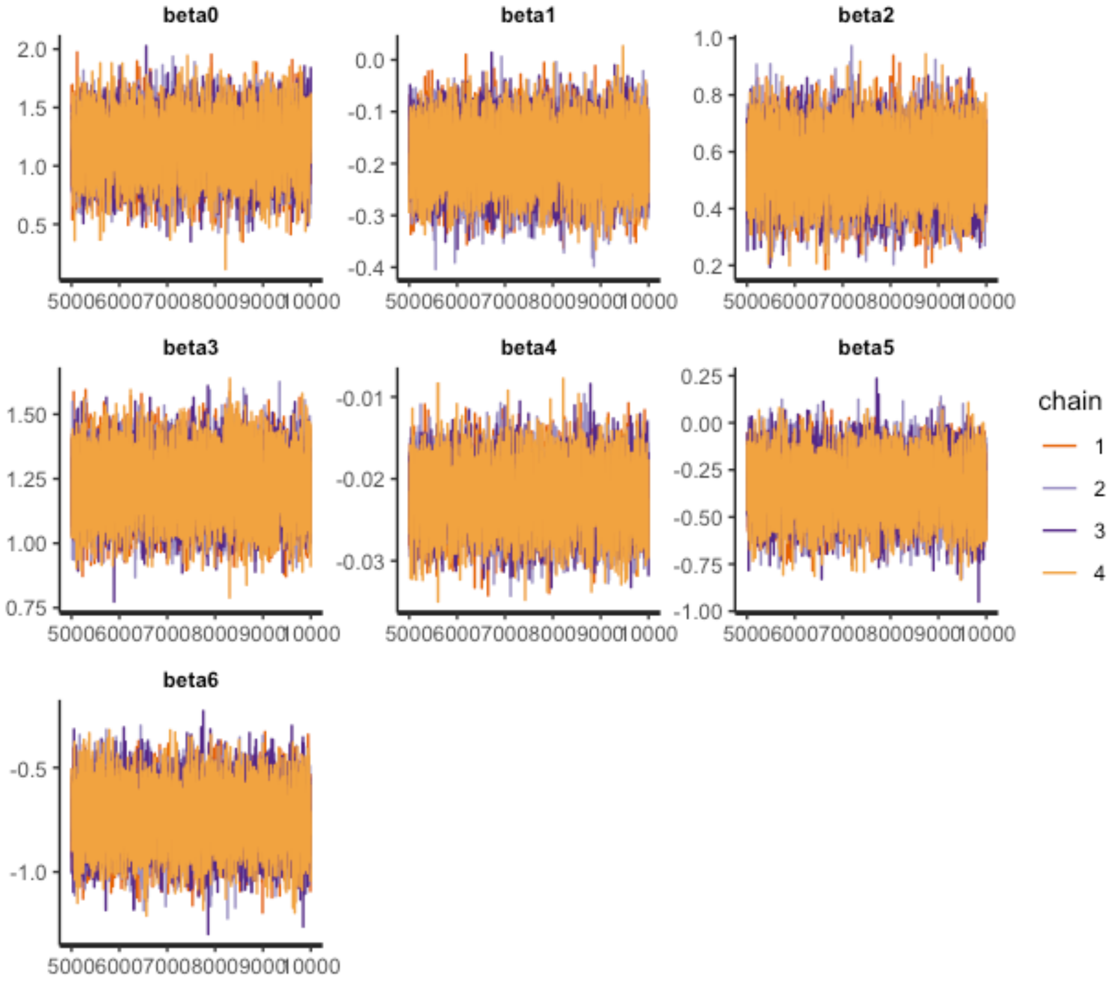

Figure 2. Trace plots for variable coefficients. The chains move around a lot and show no discernable pattern, indicating no evidence that the chains do not converge to the target distribution.

Coefficient	Gelman-Rubin statistic
β_0	1.0000502
β_1	0.9999796
β_2	1.0000520
β_3	0.9998974
β_4	1.0000530
β_5	0.9999891
β_6	1.0000180

Table 3. Gelman-Rubin statistics for variable coefficients. The Gelman-Rubin statistic is based on a comparison of within-chain and between-chain variances, and values near 1 suggest convergence

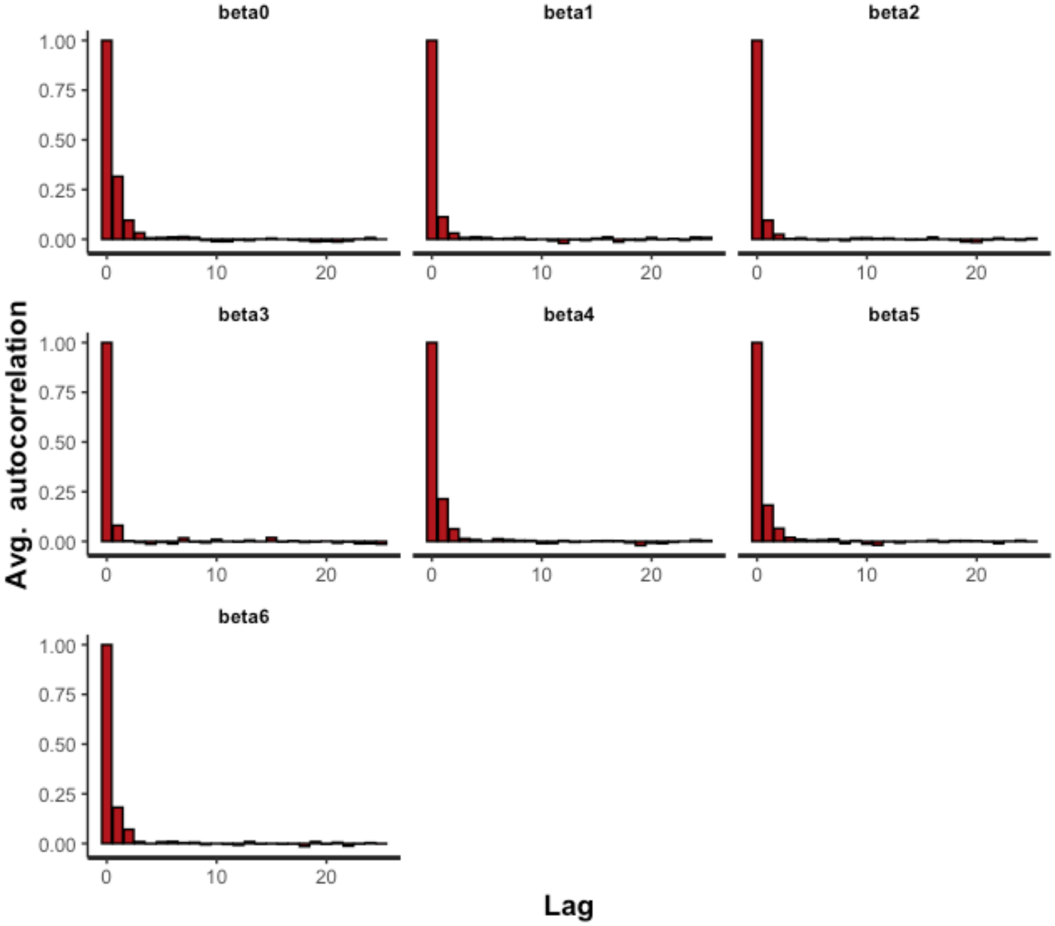

Figure 3. Autocorrelation plots for variable coefficients. The autocorrelation for each coefficient decays quickly, indicating no evidence that the chains do not converge to the target distribution.

Coefficient	Variable	Mean	Median	0.025 quantile	0.975 quantile
β_0	Intercept	1.16131652	1.16131453	0.73000932	1.59538344
β_1	NUM_KIDS	-0.18468099	-0.18513978	-0.28836680	-0.08217663
β_2	WRKLOSS	0.55375328	0.55337827	0.35442947	0.75525889
β_3	DELAY	1.22332599	1.22211855	1.01276105	1.43837431
β_4	AGE	-0.02206679	-0.02208483	-0.02874984	-0.01549604
β_5	INCOME_2	-0.34115799	-0.34281336	-0.60325047	-0.07783499
β_6	INCOME_3	-0.74028213	-0.74056456	-0.99145557	-0.48370874

Table 4. The mean, median, and 0.025 and 0.975 quantiles from the estimates for the posterior distributions of variable coefficients. Importantly, no 95% posterior credible interval (the range between the 0.025 and 0.975 quantiles) contains 0, meaning it's unlikely that the factors do not have an effect on the response.

$$\log\left(\frac{p_i}{1-p_i}\right) = 1.16 - 0.18 \cdot \texttt{NUM\_KIDS}_i + 0.55 \cdot \texttt{WRKLOSS}_i + 1.22 \cdot \texttt{DELAY}_i - 0.02 \cdot \texttt{AGE}_i - 0.34 \cdot \texttt{INCOME\_2}_i - 0.74 \cdot \texttt{INCOME\_3}_i$$

Figure 4. Our Bayesian logistic regression model. The i subscript represents the i^{th} correspondent, and p_i is the probability that the i^{th} correspondent reports feeling depressed "several days or more" during the last 7 days.

Independence Assumption

A fundamental assumption of any binomial model, including logistic regression, is that responses are independent. One way to assess the validity of that assumption in our case is to look at observed responses in sequence and count the number of switches between 1's and 0's (eg. a sequence like [0, 0, 0, 1, 1, 1] has only one switch and may be a violation of independence). We then use our posterior estimates to generate replicate data and compare the number of switches in the generated data to the number of switches in the observed data. We observe that for our model, 26% of the generated data is more extreme than the observed data, indicating no evidence to contradict the independence assumption.

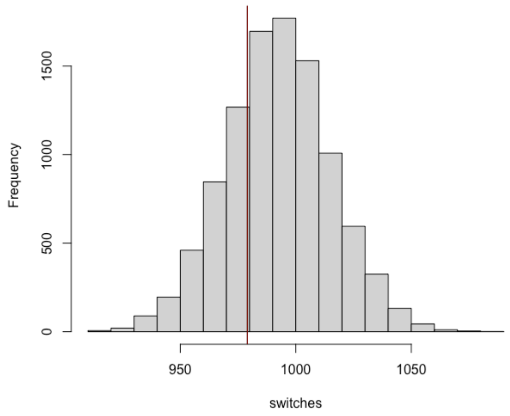

Figure 5. Verifying the independence assumption for Bernoulli outcomes. We draw 10,000 samples of replicate data and plot the number of switches between 0 and 1. The solid red line shows the number of switches in the observed data. Only 26% of replicate samples have fewer switches than the observed amount, meaning it is plausible the independence assumption is true.

Predictive Accuracy

To assess the predictive accuracy of our model, we apply it to the 17th week of data from the *Household Pulse Survey*. After applying identical preprocessing steps to the data, we estimate p_i for each respondent i using our model. If we predict p_i to be greater than 0.5, we classify that respondent as feeling depressed "several days or more", otherwise we classify them as feeling depressed "not at all." Our model achieves an accuracy of 62%, indicating that the predictive accuracy is limited.

Actual / Predicted (total 2007 responses)	**Not at all**	**Several days or more**
Not at all	480	508
Several days or more	253	766

Figure 6. A confusion matrix for cross-validation predictions.

Conclusion

For logistic regression, the interpretation of the coefficient $\beta_j, j > 0$ is that when the jth coefficient increases by one unit while holding all other variables constant, there is an a posteriori multiplicative change in the odds of "success" of e^{β_j} (in our case, "success" is feeling depressed several days or more during the last 7 days). For example, a posterior mean of 1.749 for e^{β_2} means that the expected a posteriori odds that someone reports feeling depressed several days or more during the last 7 days is 1.749 times higher for someone who has lost work compared to someone who has not, while holding all other variables constant. Estimates for e^{β_j} for $j = 0, 1, 2, 3, 4, 5, 6$ are seen in table 3.

Coefficient (β_j)	Variable	Posterior mean for e^{β_j}
β_0	Intercept	3.2749287
β_1	`NUM_KIDS`	0.8325272
β_2	`WRKLOSS`	1.7489971
β_3	`DELAY`	3.4182704
β_4	`AGE`	0.9781806
β_5	`INCOME_2`	0.7172882
β_6	`INCOME_3`	0.4809874

Table 5. Posterior estimates for e^{β_j} for $j = 0, 1, 2, 3, 4, 5, 6$. These can be interpreted as the multiplicative change in odds of reporting frequent depression when variables are increased by 1 unit (or indicator variables are switched from 0 to 1).

While considering multiplicative changes in odds may not be intuitive, we can start by noting that an increase in odds corresponds with an increase in the probability itself, and then consider the direction in which the probability of feeling depressed changes with respect to each factor based on their coefficients. We note that $\beta_1, \beta_4, \beta_5$, and β_6 are all less than 1, meaning that we expect the probability of frequent feelings of depression to be smaller for individuals who have more kids, are older, and who have more income. The coefficients β_2 and β_3 are greater than 1, meaning that we expect individuals who have lost work or have delayed medical care due to the pandemic to report more frequent feelings of depression. We can also assess the relative sizes of these effects. For example, since $\beta_3 > \beta_2$, we would estimate that delayed medical care has a greater impact on the probability of feeling depressed than losing work.

For a more intuitive understanding, we can consider a "default person" and see how modifying subsets of variables while holding others fixed changes their estimated probability of reporting feeling depressed for several days or more during the last 7 days. We let a default person be a 35 year-old individual with no kids, who has not lost work, has no delayed medical care, and an annual household income of \$60,000. Using our model, we would estimate their probability of feeling depressed several days or more to be 0.512. We then modify variables corresponding to different scenarios, seen in table 6. The largest effects occur when we consider delayed medical care (0.781), if the individual had 3 kids (0.376), or if the person was 60 years old (0.377).

Scenario	Estimated probability of feeling depressed several days or more
Default person (Age = 35, NUM_KIDS = 0, WRKLOSS = 0, DELAY = 0, INCOME_2 = 1, INCOME_3 = 0)	0.512
If the individual had 3 kids (NUM_KIDS = 3)	0.376
If the person lost work (WRKLOSS = 1)	0.646
If the person had delayed medical care (DELAY = 1)	0.781
If the person was 18 (AGE = 18)	0.604
If the person was 60 (AGE = 60)	0.377
If the person had household income less than $50,000 (INCOME_2 = 0 and INCOME_3 = 0)	0.596
If the person had household income greater than $100,000 (INCOME_2 = 0, INCOME_3 = 1)	0.413
If the person had household income less than $50,000, lost work, and has delayed medical care (INCOME_2 = 0, INCOME_3 = 0, WRKLOSS = 1, and DELAY = 1)	0.897
If the person had household income greater than $100,000, and 3 kids, and was 60 years old (INCOME_3 = 1, NUM_KIDS = 3, AGE = 60)	0.189

Table 6. How different scenarios would affect the estimated probability of reporting frequent feelings of depression compared to the baseline of a 35-year old with no kids, not lost work, no delayed medical care, and an annual household income of $60,000. The first row lists the values we select for variables and subsequent rows show how the scenario would modify the variables.

We consider a "worst-case" scenario where a person has household income less than $50,000, and has lost work and delayed medical care, for whom we estimate the probability of feeling depressed several days or more to be 0.897. For a "best-case" scenario where the individual has high income, no lost work or delayed medical care, and 3 kids, and 60 years old, we estimate the probability of feeling depressed several days or more to be 0.189.

Discussion

At first, these results may not seem surprising. Based on related work and intuition, one might guess that individuals with lower income or lost work would report feeling depressed more frequently. However, this work does shed light on a couple things that are less intuitive. First, we estimate that people are less likely to report feeling depressed if they have more children. We hypothesize that the reason for this is that while being a parent can be a source of stress, it can also combat feelings of isolation that might be stronger felt by adults living alone. Additionally, our model predicts that younger individuals are more likely to feel depressed than their older counterparts, which may be due to younger people being more economically vulnerable. Finally, we see that the biggest factor contributing to feelings of depression in our model is not loss of work, but rather delayed medical care as a result of the pandemic.

In summary, we recommend that Colorado officials direct mental health resources to individuals who are younger, have less children in the household, recent work loss, have delayed medical care, or are low income. We hope that by directing our efforts in this way, we can combat feelings of depression throughout the pandemic in Colorado residents.

Future Work

A natural extension to this work would be to consider a time component by utilizing all 17 weeks of survey data. It would be informative to know if depression is getting worse or better over time, and which factors are associated with increasing or decreasing levels of depression. Additionally, we noted that the data were unbalanced, especially with regard to race. In future work, we would apply weighting to balance the data in hopes of better capturing effects that may be present in less represented groups.

Reflection

The most challenging part of this project was learning to think in a "Bayesian" way; understanding that our estimates for model parameters are *themselves* distributions. With regard to code, I came to appreciate how simple it is to fit models using the `stan` package, especially `rstanarm`, which made fitting and comparing multiple models simple. I will note that I attempted to fit models in `rstanarm` with link functions other than logit but ran into issues I was not able to resolve. Overall, I would say that I was able to gain a lot from taking this class in addition to

Applied Regression, since I have begun to abstract some ideas for what a statistics process looks like in general, regardless of the approach.

Works Cited

Amsalem D, Dixon LB, Neria Y. The Coronavirus Disease 2019 (COVID-19) Outbreak and Mental Health: Current Risks and Recommended Actions. *JAMA Psychiatry*. Published online June 24, 2020. doi:10.1001/jamapsychiatry.2020.1730

Calliope Holingue, Elena Badillo-Goicoechea, Kira E. Riehm, Cindy B. Veldhuis, Johannes Thrul, Renee M. Johnson, M. Daniele Fallin, Frauke Kreuter, Elizabeth A. Stuart, Luther G. Kalb, "Mental distress during the COVID-19 pandemic among US adults without a pre-existing mental health condition: Findings from American trend panel survey." *Preventive Medicine*, vol. 139, 2020, https://doi.org/10.1016/j.ypmed.2020.106231.

Javed, B., Sarwer, A., Soto, E. B., & Mashwani, Z. U. (2020). The coronavirus (COVID-19) pandemic's impact on mental health. *The International journal of health planning and management*, *35*(5), 993–996. https://doi.org/10.1002/hpm.3008

Kontoangelos, K., Economou, M., & Papageorgiou, C. (2020). Mental Health Effects of COVID-19 Pandemia: A Review of Clinical and Psychological Traits. *Psychiatry investigation*, *17*(6), 491–505. https://doi.org/10.30773/pi.2020.0161

Meyer, Jacob, et al. "Changes in Physical Activity and Sedentary Behavior in Response to COVID-19 and Their Associations with Mental Health in 3052 US Adults." *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, 2020, p. 6469., doi:10.3390/ijerph17186469.

Nirmita Panchal, Rabah Kamal, and Apr 2020. "The Implications of COVID-19 for Mental Health and Substance Use." *KFF*, 21 Aug. 2020, www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/.

Purtle J (2012) “Racial and ethnic disparities in post-disaster mental health: examining the evidence through a lens of social justice.” *Wash Lee J Civ Rights Soc Justice* 19:31.

US Census Bureau. *2020 Household Pulse Survey Week 16*, Oct. 2020, www.census.gov/programs-surveys/household-pulse-survey/datasets.html.

Zhou, Yanmengqian, et al. “Mental Health and Its Predictors during the Early Months of the COVID-19 Pandemic Experience in the United States.” *International Journal of Environmental Research and Public Health*, vol. 17, no. 17, 2020, p. 6315., doi:10.3390/ijerph17176315.

Appendix

Variable	Definition in Household Pulse Survey	Recoding in this project
GENDER	EGENDER (Gender) 1) Male 2) Female	1 if EGENDER = 2 0 if EGENDER = 1
AGE	TBIRTH_YEAR (Year of birth) 1932-2002	2020 – TBIRTH_YEAR
RACE	RRACE (Race) 1) White, Alone 2) Black, Alone 3) Asian, Alone 4) Any other race alone, or race in combination	1 if RRACE = 2, 3, or 4 or if RHISPANIC = 2 0 if RRACE = 1 and RHISPANIC = 1
	RHISPANIC (Hispanic origin) 1) No, not of Hispanic, Latino, or Spanish origin 2) Yes, of Hispanic, Latino, or Spanish origin	
EDUC	EEDUC (Educational attainment) 1) Less than high school 2) Some high school 3) High school graduate or equivalent (for example GED) 4) Some college, but degree not received or is in progress 5) Associate's degree (for example AA, AS) 6) Bachelor's degree (for example BA, BS, AB) 7) Graduate degree (for example master's, professional, doctorate)	1 if EEDUC = 5, 6, or 7 0 if EEDUC = 1, 2, 3, or 4
MS	MS (Marital status) 1) Now married 2) Widowed 3) Divorced 4) Separated 5) Never married -99) Question seen but category not selected -88) Missing / Did not report	1 if MS = 1 0 if MS = 2, 3, 4, or 5
NUM_KID	THHLD_NUMKID (Total number of people under 18-years-old in household) (0-40) number of people under 18 (whole number)	No change
WRKLOSS	WRKLOSS (Recent household job loss) 1) Yes 2) No -99) Question seen but category not selected -88) Missing / Did not report	1 if WRKLOSS = 1 0 if WRKLOSS = 2
DELAY	DELAY (Delayed medical care in last 4 weeks due to pandemic)	1 if DELAY = 1 0 if DELAY = 2

	1) Yes 2) No -99) Question seen but category not selected -88) Missing / Did not report	
RENT	TENURE (Housing owned or rented) 1) Owned free and clear? 2) Owned with a mortgage or loan (including home equitly loans)? 3) Rented? 4) Occupied without payment of rent? -99) Question seen but category not selected -88) Missing / Did not report	1 if TENURE = 2 or 3 0 if TENURE = 1 or 4
INCOME_1	INCOME (Total household income (before taxes)) 1) Less than $25,000 2) $25,000 - $34,999 3) $35,000 - $49,999 4) $50,000 - $74,999 5) $75,000 - $99,999 6) $100,000 - $149,999 7) $150,000 - $199,999 8) $200,000 and above -99) Question seen but category not selected -88) Missing / Did not report	1 if INCOME = 1, 2, or 3 0 if INCOME = 4, 5, 6, 7, or 8
INCOME_2		1 if INCOME = 4 or 5 0 if INCOME = 1, 2, 3, 6, 7, or 8
INCOME_3		1 if INCOME = 6, 7, or 8 0 if INCOME = 1, 2, 3, 4, or 5
DOWN	DOWN (Frequency of feeling depressed over previous 7 days) 1) Not at all 2) Several days 3) More than half the days 4) Nearly every day -99) Question seen but category not selected -88) Missing / Did not report	1 if DOWN = 2, 3, or 4 0 if DOWN = 1

Table 7. Factors considered in our analysis with how they were defined in the Census Household Pulse Survey, and how we defined them in our analysis.

R Code

```
getwd()
setwd('/Users/Nick/Desktop')

################################################################
################################################################
# Read in and clean up data
################################################################
################################################################

data = read.csv('pulse2020_puf_16.csv')

# Narrow down to only colorado
data = data[data$EST_ST == 8,]
data = data[,c('TBIRTH_YEAR', 'EGENDER', 'RRACE', 'RHISPANIC',
               'EEDUC', 'MS', 'THHLD_NUMKID', 'WRKLOSS',
               'DELAY', 'TENURE', 'INCOME', 'DOWN')]

data['Age'] = 2020 - data[,'TBIRTH_YEAR']
data = data[,colnames(data) != 'TBIRTH_YEAR']

for(column in colnames(data)){
  data[,column][data[,column] == -88] = NA
  data[,column][data[,column] == -99] = NA

  #if(column != 'Age'& column != 'THHLD_NUMKID'){
  #  data[,column] = factor(data[,column])
  #}
}

# Drop missing data
data = na.omit(data)

# Recode variables

# Set DOWN to 0 for 'not at all', otherwise 1
data$DOWN = (data$DOWN != 1) + 0

# Set RACE to 0 for white, 1 for non-white
data$RACE = (data$RRACE != 1 | data$RHISPANIC != 1) + 0
data = data[,!(colnames(data) %in% c('RRACE', 'RHISPANIC'))]

# Set EDUC to 0 for not college degree, else 1
data$EDUC = (data$EEDUC > 4) + 0
data = data[,!(colnames(data) %in% c('EEDUC'))]
```

```
# Set MS to 0 for single, else 1
data$MS = (data$MS == 1) + 0

# Set RENT to 1 if rent or mortgage, otherwise 0
data$RENT = ((data$TENURE == 2) | (data$TENURE == 3)) + 0
data = data[,!(colnames(data) %in% c('TENURE'))]

# What should the income threshold be?
for (eq in 1:8){
  a = dim(data[(data$DOWN == 1 & data$INCOME ==
eq),])[1]/dim(data[(data$DOWN == 0 & data$INCOME == eq),])[1]
  print(eq)
  print(a)
  print('')
}

# It looks like cutoffs should be 3 - 4 ($50,000)
# and 5 - 6 ($100,000)
data$INCOME_2 = (data$INCOME > 3 & data$INCOME <= 5) + 0
data$INCOME_3 = (data$INCOME > 5) + 0

# Replace the income variable with one that has 3 levels that we
can use
# For EDA. We'll drop before building the model
data$INCOME = 0
data = within(data, INCOME[INCOME_2 == 1] <- 1)
data = within(data, INCOME[INCOME_3 == 1] <- 2)

# Rename the gender column
colnames(data)[1] = 'GENDER'
data$GENDER = data$GENDER - 1 # Now female is 1

# Rename the number of children column
colnames(data)[3] = "NUM_KIDS"

# Recode delayed medical care to be 1 if delayed medical care,
else 0
data$DELAY = data$DELAY - 1
data$DELAY = (data$DELAY == 0) + 0

# Recode work loss to be 0 if no job loss
data$WRKLOSS = data$WRKLOSS - 1 # Now female is 1
data$WRKLOSS = (data$WRKLOSS == 0) + 0

##################################################################
##################################################################
```

```
# Data exploration
################################################################
################################################################

##################
# Balance of data
##################

# DOWN
dim(data[data$DOWN ==1,])/1992

# GENDER
dim(data[data$GENDER ==1,])/1992

# MS

# NUM_KIDS

# WRKLOSS

# DELAY

# INCOME

# Age

# RACE

# EDUC

# RENT



# Tables for categorical data
####################
# Gender (1 is male)
####################
a = table(data$DOWN, data$GENDER)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('Male', 'Female'), col =
c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topleft"),
        main = 'Frequency of feeling depressed over previous 7
days', ylab = 'Count')

# More females are depressed than not, men are about even

####################
# Marriage status (1 is married)
####################
a = table(data$DOWN, data$MS)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('Single', 'Married'),
col = c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topleft"),
        main = 'Frequency of feeling depressed over previous 7
days', ylab = 'Count')

# Single people tend to be more depressed than not

####################
# Work loss **
####################
a = table(data$DOWN, data$WRKLOSS)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('No', 'Yes'), col =
c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topright"), xlab = "Recent household job
loss", cex.names = 1.5, cex.lab = 1.5)

# Among the people who have lost work, much more of them are
depressed than not

####################
# Delayed medical care **
####################
a = table(data$DOWN, data$DELAY)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('No', 'Yes'), col =
c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topright"), xlab = "Delayed medical care
during last 4 weeks due to pandemic", cex.names = 1.5, cex.lab =
1.5)

# People with delayed medical care tend to be a lot more
depresed than not, and visa-versa

####################
# Race
####################
a = table(data$DOWN, data$RACE)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('White', 'Non-white'),
col = c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topright"),
        main = 'Frequency of feeling depressed over previous 7
days', ylab = 'Count')

# Race is close, but imbalanced because most respondents were
white

####################
# Education
####################
a = table(data$DOWN, data$EDUC)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('Non-college educated',
'College educated'), col = c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topleft"),
        main = 'Frequency of feeling depressed over previous 7
days', ylab = 'Count')

# Non college educated tend to feel down more than not, but
again it's imbalanced because
# most respondents were college educated

####################
# Rent/Mortgage
####################
a = table(data$DOWN, data$RENT)
```

```
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('No rent or mortgage',
'Pay rent or mortgage'), col = c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topleft"),
        main = 'Frequency of feeling depressed over previous 7
days', ylab = 'Count')

# People who pay rent or mortgage tend to be more depressed than
not and visa-versa

####################
# Income **
####################
a = table(data$DOWN, data$INCOME)
a/margin.table(a)
prop.table(a, 2)

barplot(a, beside = TRUE, names.arg = c('<$50,000', '$50,000-
$100,000', '>$100,000'), col = c("darkblue", "red"),
        legend = c('Not at all', 'Several days or more'),
args.legend = list(x = "topleft"), xlab = "Household income",
cex.names = 1.5, cex.lab = 1.5)

####################
# Box plots for numeric variables
####################
####################
# Age **
####################
#boxplot(data$Age~data$DOWN, ylab = 'Age', xlab = 'Frequency of
feeling depressed over previous 7 days',
#        names = c('Not at all', 'Several days or more'), col =
c("darkblue", "red"))

boxplot(data$Age~data$DOWN, ylab = 'Age',
        names = c('Not at all', 'Several days or more'), col =
c("darkblue", "red"), xlab = "", cex.names = 1.5, cex.lab = 1.5)

a = table(data$DOWN, data$Age)
plot(c(0, 1, 2, 3, 4, 5), prop.table(a, 2)[1,])
lmod = lm(prop.table(a, 2)[1,]~c(0, 1, 2, 3, 4, 5))
summary(lmod)

####################
```

```
# Number of children
####################
boxplot(data$NUM_KIDS~data$DOWN, ylab = 'Number of children',
xlab = 'Frequency of feeling depressed over previous 7 days',
        names = c('Not at all', 'Several days or more'), col =
c("darkblue", "red"))

a = table(data$DOWN, data$NUM_KIDS)
plot(c(0, 1, 2, 3, 4, 5), prop.table(a, 2)[1,])
lmod = lm(prop.table(a, 2)[1,]~c(0, 1, 2, 3, 4, 5))
summary(lmod)
################################################################
################################################################
# Backward selection with frequentist to remove some variables
using rstanarm
################################################################
################################################################

####################
# Full model
####################
data = data[,!(colnames(data) %in% 'INCOME')]

glmmod = glm(data2$DOWN ~., data = data, family = binomial)
summary(glmmod)

# Remove race
data2 = data[,!(colnames(data) %in% c('RACE'))]
glmmod = glm(data2$DOWN ~., data = data2, family = binomial)
summary(glmmod)

# Remove Rent
data2 = data2[,!(colnames(data2) %in% c('RENT'))]
glmmod = glm(data2$DOWN ~., data = data2, family = binomial)
summary(glmmod)

# Check with a chi-square test
chisq.test(data$DOWN, data$RENT) # 0.08595, so maybe it's
explained by something else

# Remove education
data2 = data2[,!(colnames(data2) %in% c('EDUC'))]
glmmod = glm(data2$DOWN ~., data = data2, family = binomial)
summary(glmmod)

# Now all p-values are significant at the 0.10 level and switch
to Bayes
```

```
################################################################
################################################################
# now try fitting some models with stanarm
################################################################
################################################################
data = data2

# full model
glmod_logit = stan_glm(DOWN ~ ., data = data2,
                       family = binomial(link = "logit"),
                       prior = normal(0, sqrt(1000)),
                       prior_intercept = normal(0, sqrt(1000)),
                       iter = 10000, chains = 4, seed = 101)

summary(glmod_logit)

# Try fitting models without variables that have a small impact:
# GENDER, MS, NUM_KIDS, and Age

##############
# No gender
##############
glmod_logit_nogender = stan_glm(DOWN ~ .-GENDER, data = data2,
                       family = binomial(link = "logit"),
                       prior = normal(0, sqrt(1000)),
                       prior_intercept = normal(0, sqrt(1000)),
                       iter = 10000, chains = 4, seed = 101)

summary(glmod_logit_nogender)

##############
# No MS
##############
glmod_logit_noms = stan_glm(DOWN ~ .-MS, data = data2,
                                family = binomial(link =
"logit"),
                                prior = normal(0, sqrt(1000)),
                                prior_intercept = normal(0,
sqrt(1000)),
                                iter = 10000, chains = 4, seed =
101)

summary(glmod_logit_noms)

##############
# No NUM_KIDS
```

```
##############
glmod_logit_nokids = stan_glm(DOWN ~ .-NUM_KIDS, data = data2,
                            family = binomial(link = "logit"),
                            prior = normal(0, sqrt(1000)),
                            prior_intercept = normal(0,
sqrt(1000)),
                            iter = 10000, chains = 4, seed =
101)

summary(glmod_logit_nokids)

##############
# No Age
##############
glmod_logit_noage = stan_glm(DOWN ~ .-Age, data = data2,
                              family = binomial(link = "logit"),
                              prior = normal(0, sqrt(1000)),
                              prior_intercept = normal(0,
sqrt(1000)),
                              iter = 10000, chains = 4, seed =
101)

summary(glmod_logit_noage)

####### Compare WAIC and LOOIC
waic_full = waic(glmod_logit)
loo_full = loo(glmod_logit)
waic_nogender = waic(glmod_logit_nogender)
loo_nogender = loo(glmod_logit_nogender)
waic_noms = waic(glmod_logit_noms)
loo_noms = loo(glmod_logit_noms)
waic_nokids = waic(glmod_logit_nokids)
loo_nokids = loo(glmod_logit_nokids)
waic_noage = waic(glmod_logit_noage)
loo_noage = loo(glmod_logit_noage)

loo_compare(waic_full, waic_nogender, waic_noms, waic_nokids,
waic_noage)
loo_compare(loo_full, loo_nogender, loo_noms, loo_nokids,
loo_noage)

# Could remove gender, noms
glmod_logit_nogender_noms = stan_glm(DOWN ~ .-(GENDER + MS),
data = data2,
                             family = binomial(link = "logit"),
                             prior = normal(0, sqrt(1000)),
                             prior_intercept = normal(0, 
sqrt(1000)),
                             iter = 10000, chains = 4, seed = 
101)

# Could remove gender, kids
glmod_logit_nogender_nokids = stan_glm(DOWN ~ .-(GENDER + 
NUM_KIDS), data = data2,
                                     family = binomial(link = 
"logit"),
                                     prior = normal(0, 
sqrt(1000)),
                                     prior_intercept = normal(0, 
sqrt(1000)),
                                     iter = 10000, chains = 4, 
seed = 101)


# Could remove ms, kids
glmod_logit_noms_nokids = stan_glm(DOWN ~ .-(NUM_KIDS + MS), 
data = data2,
                                     family = binomial(link = 
"logit"),
                                     prior = normal(0, 
sqrt(1000)),
                                     prior_intercept = normal(0, 
sqrt(1000)),
                                     iter = 10000, chains = 4, 
seed = 101)

# Remove ms, gender, kids
glmod_logit_nogender_noms_nokids = stan_glm(DOWN ~ .-(GENDER + 
NUM_KIDS + MS), data = data2,
                                   family = binomial(link = 
"logit"),
                                   prior = normal(0, 
sqrt(1000)),
                                   prior_intercept = normal(0, 
sqrt(1000)),
                                   iter = 10000, chains = 4, 
seed = 101)

waic_nogender_noms = waic(glmod_logit_nogender_noms)
waic_nogender_nokids = waic(glmod_logit_nogender_nokids)
waic_noms_nokids = waic(glmod_logit_noms_nokids)
waic_nogender_noms_nokids = 
waic(glmod_logit_nogender_noms_nokids)
```

```
looic_nogender_noms = loo(glmod_logit_nogender_noms)
looic_nogender_nokids = loo(glmod_logit_nogender_nokids)
looic_noms_nokids = loo(glmod_logit_noms_nokids)
looic_nogender_noms_nokids =
loo(glmod_logit_nogender_noms_nokids)

loo_compare(waic_full, waic_nogender_noms, waic_nogender_nokids,
waic_noms_nokids)
loo_compare(loo_full, looic_nogender_noms,
looic_nogender_nokids, looic_noms_nokids)

models = c('Full', 'Without GENDER', 'Without MS', 'Without
NUM_KIDS', 'Without Age',
           'Without GENDER or MS', 'Without GENDER or NUM_KIDS',
'Without MS or NUM_KIDS',
           'Without GENDER or MS or NUM_KIDS')

waics = c(waic_full$waic, waic_nogender$waic, waic_noms$waic,
waic_nokids$waic, waic_noage$waic, waic_nogender_noms$waic,
          waic_nogender_nokids$waic, waic_noms_nokids$waic,
waic_nogender_noms_nokids$waic)

loos = c(loo_full$looic, loo_nogender$looic, loo_noms$looic,
loo_nokids$looic, loo_noage$looic, looic_nogender_noms$looic,
          looic_nogender_nokids$looic, looic_noms_nokids$looic,
looic_nogender_noms_nokids$looic)

results = data.frame(models, waics, loos)

loo_compare(waic_full, waic_nogender, waic_noms, waic_nokids,
waic_noage, waic_nogender_noms, waic_nogender_nokids,
waic_noms_nokids, waic_nogender_noms_nokids)
loo_compare(loo_full, loo_nogender, loo_noms, loo_nokids,
loo_noage, looic_nogender_noms, looic_nogender_nokids,
looic_noms_nokids, looic_nogender_noms_nokids)

# Remove gender and ms from the model because they're within one
standard deviation
colnames(data)
# Final model will have NUM_KIDS, WRKLOSS, DELAY, Age, INCOME_2,
and INCOME_3
summary(glmod_logit_nogender_noms)

# Number of kids -> Less likely to be depressed (Maybe having
more people in the household means less feeling of isolation)
# Loss of work -> More likely to be depressed
# Delayed medical care -> More likely to be depressed
# Younger -> More likely to be depressed
# Higher income -> Less likely to be depressed

################################################################
################################################################
# Try different priors and compare WAIC, LOOIC
################################################################
################################################################

glmod_logit_nogender_noms = stan_glm(DOWN ~ .-(GENDER + MS),
data = data2,
                                     family = binomial(link =
"logit"),
                                     prior = normal(0,
sqrt(1000)),
                                     prior_intercept = normal(0,
sqrt(1000)),
                                     iter = 10000, chains = 4,
seed = 101)

glmod_probit_nogender_noms = stan_glm(DOWN ~ .-(GENDER + MS),
data = data2,
                                     family = binomial(link =
"probit"),
                                     prior = normal(0,
sqrt(1000)),
                                     prior_intercept = normal(0,
sqrt(1000)),
                                     iter = 10000, chains = 4,
seed = 101, init_r = 0.5)

glmod_cloglog_nogender_noms = stan_glm(DOWN ~ .-(GENDER + MS),
data = data2,
                                     family = binomial(link =
"cloglog"),
                                     prior = normal(0,
sqrt(1000)),
                                     prior_intercept = normal(0,
sqrt(1000)),
                                     iter = 10000, chains = 4,
seed = 101, init_r = 0.5)

waic_logit = waic(glmod_logit_nogender_noms_nokids)
waic_probit = waic(glmod_probit_nogender_noms_nokids)
waic_cloglog = waic(glmod_cloglog_nogender_noms_nokids)
```

```
looic_logit = loo(glmod_logit_nogender_noms)
looic_probit = loo(glmod_probit_nogender_noms)
looic_cloglog = loo(glmod_cloglog_nogender_noms)

################################################################
################################################################
# Fitting the model and checking covergence
################################################################
################################################################

stanmod = "
data {
  int<lower=1> n; // number of observations
  int y[n];       // indicator of senility
  vector[n] num_kids;
  vector[n] work_loss;
  vector[n] delayed_care;
  vector[n] age;
  vector[n] income_2;
  vector[n] income_3;
}
parameters {
  real beta0;
  real beta1; // num_kids
  real beta2; // work_loss
  real beta3; // delayed_care
  real beta4; // age
  real beta5; // income_2
  real beta6; // income_3
}
transformed parameters {
  vector[n] mu;           // mean of observations
  for(i in 1:n){
    mu[i] = beta0 + beta1*num_kids[i] + beta2*work_loss[i] +
beta3*delayed_care[i] + beta4*age[i] + beta5*income_2[i] +
beta6*income_3[i];
  }
}
model {
  // prior distributions
  beta0 ~ normal(0, sqrt(1000));
  beta1 ~ normal(0, sqrt(1000));
  beta2 ~ normal(0, sqrt(1000));
  beta3 ~ normal(0, sqrt(1000));
  beta4 ~ normal(0, sqrt(1000));
  beta5 ~ normal(0, sqrt(1000));
  beta6 ~ normal(0, sqrt(1000));
  // data distribution
  // data distribution
  for(i in 1:n){
    y[i] ~ bernoulli_logit(mu[i]);
  }
}
generated quantities {
  real exp_beta0;
  real exp_beta1;
  real exp_beta2;
  real exp_beta3;
  real exp_beta4;
  real exp_beta5;
  real exp_beta6;

  vector[n] log_lik;  // log likelihood of data

  exp_beta0 = exp(beta0);
  exp_beta1 = exp(beta1);
  exp_beta2 = exp(beta2);
  exp_beta3 = exp(beta3);
  exp_beta4 = exp(beta4);
  exp_beta5 = exp(beta5);
  exp_beta6 = exp(beta6);

  for (i in 1:n) log_lik[i] = bernoulli_logit_lpmf(y[i]|mu[i]);
}
"

stan_dat = list(n = length(data$DOWN), y = data$DOWN, num_kids =
data$NUM_KIDS, work_loss = data$WRKLOSS,
                delayed_care = data$DELAY, age = data$Age,
income_2 = data$INCOME_2,
                income_3 = data$INCOME_3)

stan_fit = stan(model_code = stanmod, data = stan_dat, iter =
100000)

coeffs = c("beta0", "beta1", "beta2", "beta3", "beta4", "beta5",
"beta6")

summary(stan_fit)$summary[coeffs, c("mean", "50%", "2.5%",
"97.5%")]

# Gelman-Rubin statistic
summary(stan_fit)$summary[coeffs,"Rhat"]
```

```
# check convergence with trace plots
stan_trace(stan_fit, coeffs)

# Check the ACF of draws
stan_ac(stan_fit, coeffs)

log_lik = extract_log_lik(stan_fit, merge_chains = FALSE)
r_eff = exp(relative_eff(log_lik))

waic(log_lik) # want models with smaller waic
# 2470.2

loo(log_lik, r_eff = r_eff) # want models with a smaller looic
# 2470.2

#####################
# Posterior densities
#####################
stan_dens(stan_fit, par = c("beta0", "beta1", "beta2", "beta3",
"beta4", "beta5", "beta6"),
          separate_chains = TRUE)

################################################################
################################################################
# Posterior checks
################################################################
################################################################

library(bayesplot)

B = 1e4 #10000 samples

sumy = sum(data$DOWN)
n = length(data$DOWN)
y = data$DOWN
# sample from posterior distribution of theta
theta = rbeta(B, sumy + 1, n - sumy + 1)

# number of switches between 0 and 1 using rle
# (run length encoding) function
nswitch = function(x) length(rle(x)$values) - 1

# generate yrep
yrep = matrix(0, nrow = B, ncol = n)
for (i in 1:B) {
  # sample n bernoulli trials with success probability theta[i]
  yrep[i, ] = rbinom(n, size = 1, prob = theta[i])
}

nswitch_y = nswitch(y)
nswitch_yrep = apply(yrep, 1, nswitch)

# posterior predictive probability
(sum(nswitch_yrep >= nswitch_y) + 1)/(B + 1)

# histogram of number of switches for yrep
# compared to observed value
hist(nswitch_yrep, xlab = "switches", main = "")
abline(v = nswitch_y, col = "darkred")

ppc_stat(y, yrep, stat = nswitch)

# The null hypothesis that it's just random chance
# is random chance, there is no lack of fit

# Bayesian p-value is the "probability that the replicated
# data could be more extreme than the observed data

################################################################
################################################################
# Cross-validation on week 17
################################################################
################################################################

val = read.csv('pulse2020_puf_17.csv')

# Narrow down to only colorado
val = val[val$EST_ST == 8,]
val = val[,c('TBIRTH_YEAR', 'THHLD_NUMKID', 'WRKLOSS',
               'DELAY', 'INCOME', 'DOWN')]

val['Age'] = 2020 - val[,'TBIRTH_YEAR']
val = val[,colnames(val) != 'TBIRTH_YEAR']

for(column in colnames(val)){
  val[,column][val[,column] == -88] = NA
  val[,column][val[,column] == -99] = NA
}

# Drop missing data
val = na.omit(val)

# Recode variables
# Set DOWN to 0 for 'not at all', otherwise 1
val$DOWN = (val$DOWN != 1) + 0

# It looks like cutoffs should be 3 - 4 ($50,000)
# and 5 - 6 ($100,000)
val$INCOME_2 = (val$INCOME > 3 & val$INCOME <= 5) + 0
val$INCOME_3 = (val$INCOME > 5) + 0

# Rename the number of children column
colnames(val)[1] = "NUM_KIDS"

# Recode delayed medical care to be 1 if delayed medical care,
else 0
val$DELAY = val$DELAY - 1
val$DELAY = (val$DELAY == 0) + 0

# Recode work loss to be 0 if no job loss
val$WRKLOSS = val$WRKLOSS - 1 # Now female is 1
val$WRKLOSS = (val$WRKLOSS == 0) + 0

val = val[,!(colnames(val) %in% 'INCOME')]

a = summary(stan_fit)$summary[coeffs, c("mean")]

summ = exp(a['beta0'] + a['beta1']*val$NUM_KIDS +
a['beta2']*val$WRKLOSS + a['beta3']*val$DELAY +
             a['beta4']*val$Age + a['beta5']*val$INCOME_2 +
a['beta4']*val$INCOME_3)

b = summ/(1 + summ)

r = c()
for(i in b){
  if(i > 0.5){
    r = c(r, 1)
  }
  else{
    r = c(r, 0)
  }
}
sum(r == val$DOWN)/length(val$DOWN)

# 62% accurate on a validation set

# confusion matrix
confusionMatrix(as.factor(val$DOWN), as.factor(r))
```

```
#################################################################
#################################################################
# Interpreting coefficients
#################################################################
#################################################################

########
# Estimates for exp(beta_j) for the coefficients
########
summary(stan_fit)$summary[c("exp_beta0", "exp_beta1",
"exp_beta2", "exp_beta3",
                            "exp_beta4", "exp_beta5",
"exp_beta6"), c("mean")]


# Default person

kids = 0
work_loss = 0
delayed_care = 0
age = 35
income2 = 1
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.512

# If that person had 3 kids

kids = 3
work_loss = 0
delayed_care = 0
age = 35
income2 = 1
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.376
```

```
# If the person lost work

kids = 0
work_loss = 1
delayed_care = 0
age = 35
income2 = 1
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.646

# If the person had delayed medical care

kids = 0
work_loss = 0
delayed_care = 1
age = 35
income2 = 1
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.781

# If the person was 18 years old

kids = 0
work_loss = 0
delayed_care = 0
age = 18
income2 = 1
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.604
```

```
# If the person was 60 years old

kids = 0
work_loss = 0
delayed_care = 0
age = 60
income2 = 1
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.377

# If the person made less than $50,000

kids = 0
work_loss = 0
delayed_care = 0
age = 35
income2 = 0
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) # 0.596

# If the person was over $100,000

kids = 0
work_loss = 0
delayed_care = 0
age = 35
income2 = 0
income3 = 1

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)
```

```
print(summ/(1 + summ)) # 0.413

# Worst case scenario: low income, lost work and delayed medical
care

kids = 0
work_loss = 1
delayed_care = 1
age = 35
income2 = 0
income3 = 0

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) #0.897

# Best case scenario: high income, no lost work or delayed
medical care, 3 kids, and 60 years old

kids = 3
work_loss = 0
delayed_care = 0
age = 60
income2 = 0
income3 = 1

summ = exp(a['beta0'] + a['beta1']*kids + a['beta2']*work_loss +
a['beta3']*delayed_care +
             a['beta4']*age + a['beta5']*income2 +
a['beta6']*income3)

print(summ/(1 + summ)) #0.189
```