Exploring the Relationship between Macroeconomic Indicators and Public Survery Responses to Economic Questions

Daniel Fanelli (daniel.fanelli@spsmail.cuny.edu), Christopher Fenton (christopher.fenton@spsmail.cuny.edu), James Hamski (james.hamski@spsmail.cuny.edu), Youqing Xiang (youqing.xiang@spsmail.cuny.edu)

7/21/2016

Abstract

The link between consumer sentiment and macroeconomic conditions may be studied by public surveys. We attempt to answer the question "does public sentiment about the economy reflect measured macroeconomic indicators?" and show that demographic factors such as income and education level attained are better predictors of a survey respondent's assessment of the economy than official measures of economic activity.

Keywords
survey analyis, economics, linear regression, labor markets, household finances

1 Introduction

It was noted by Pew Research Center in 2014 that economic pessimism in the post-Great Recession period (since 2008) is prevalent despite improving economic indicators such as unemployement rate and stock market value Pew, 2014. In this paper, we use the Federal Reserve Bank of New York's Survey of Consumer Expectations (SCE) to investigate the relationship between public perception of employement and financial status in the context of macroeconomic indicators such as unemployement as recorded by the Federal Reserve Bank of St. Louis's FRED database. We attempt to answer the question "does public sentiment about the economy reflect measured macroeconomic indicators?" and show that demographicfactors such as income and education level attained are better predictors of a survey respondent's assessment of the economy than official measures of economic activity.

1.1 The Federal Reserve Bank of New York's Suvery of Consumer Expectations

The Federal Reserve Bank of New York's (NYFRB) Suvery of Consumer Expectations (SCE) is designed to elicit consumers' expectations for a wide range of household-level and aggregate economic and financial conditions. Their 'microdata' release contains survey response data from June 2013 through September 2015. The microdata contains 37473 survey responses, however not all questions contain a response from all respondents, either due to missing data or the introduction of a question after the launch of the survey.

The goal of the SCE is to measure an individual's beliefs about the likelihood of future outcomes and how certain or uncertain the respondent is about future economic and household financial conditions.

According to the FRBNY: *The SCE is a nationally representative, internet-based survey of a rotating panel of approximately 1,200 household heads. Respondents participate in the panel for up to twelve months, with a roughly equal number rotating in and out of the panel each month. Unlike comparable surveys based on repeated cross-sections with a different set of respondents in each wave, our panel allows us to observe the changes in expectations and behavior of the same individuals over time.* (2013)

The SCE includes information about the age, location, race, education level, and income of the respondent. The microdata includes a column of weights that allow for the the weighted analysis of the data in such a way that reflects the demographics of the United States as a whole, as the survey respondents skew more educated and higher income than the national averages. The models built in this paper utilize these weights.

This paper utilizes several of these demographic factors as predictor variables (explained below). However, the authors decided to not utilize race as a factor for further analysis due to potential errors in variable coding introduced during data cleaning and the belief that other factors such as income can better capture underlying factors of economic belief.

1.2 FRED Data

The Federal Reserve Bank of St. Lous (FRBSL) maintains the FRED economic data repository. It represents the most easily obtained data source for common macroeconomic indicators such as unemployement rate, inflation, and aggregate market data. In this paper we obtained seven macroeconomic indicators from the FRED data repository and joined their monthly value to the month of each survey response. In this way, we can know the value for seven different macroeconomic measures for the same month a survey response was entered. The following data are used as predictor variables:
1) Unemployement Rate (UNRATE)
2) Total Nonfarm Payroll (PAYEMS)
3) 4-Week Moving Average of Initial Claims (IC4WSA)
4) KC Fed Labor Market Conditions Index, Momentum Indicator (FRBKCLMCIM)
5) Change in Labor Market Conditions Index (FRBLMCI)
6) Consumer Price Index (CPIAUCSL)
7) S&P 500 (SP500)

2 Literature Review

There has been a substantial amount of research analyzing the relationship between consumer sentiment and indicators of economic performance. Much of this research utlizes the University of Michigan Consumer Confidence Index, which contains two questions which ask respondents to assess present economic conditions. Evidence from regression analysis suggests that measures of consumer confidence—taken alone—have important predictive power for quarterly consumer expenditure growth (Ludvigson, 2004). This research provides a framework for us to build on. At the empirical level, when examining consumer sentiment survey data, survey micro data offers more depth than aggregate/index data (Souleles, 2004). This is especially important in the context of demographic distinctions, as survey microdata preserves these dinstinctions that are often lost when aggregating.

Consumer sentiment has been linked to several external factors. News media influences consumer sentiment through the tone, volume, and frequency of the delivery (Doms and Morin, 2004). Further, this reporting can diverge from fundamental indicators, which may lead to a gap between sentiment and said indicators.

There is also evidence that consumer sentiment is sensitive to factors that are more apparent to them, such as home price changes in their local market or whether or not they have experienced unemployment personally (Kuchler and Zafar, 2015).

Our analysis attempts to build on this body of research by exploring whether demographic categorizations are a better predictor of consumer sentiment than macroeconomic indicators. Souleles established that demographics had a signficant effect on the individual specific error of their sentiment against that individual's outcome.

3 Methodology

3.1 Test vs. Train

The dataset was partitioned into an 80% data frame for training and a 20% dataset for testing. This resulted in 29823 responses in the training dataset and 7650 responses in the test dataset.

3.2 Response Variables

Three response variables were selected from the SCE as indicators of public sentiment and expectations about the economy: (1) Probability of decrease in earnings 12 months from now, (2) Financially better or worse off 12 months from now, (3) Percent chance 12 months from now unemployement higher. The SCE features several additional questions that would have been appropriate to explore as response variables, however these were chosen beause they are asking about different dimensions of the respondent's views about the economy - (1) and (2) are asking about personal economic outcomes, while (3) asks about the economy as a whole. (1) and (3) are asking for a continuous numeric response, while (2) is asking for a categorical response along the lines of "Much better off" or "Worse off". This allows us to utilize linear regression techniques used for continuous response variables and multinomial linear regression.

3.2.1 Response Variable 1: Probability of decrease in earnings 12 months from now

This is a continuous response variable, generated by asking the respondent to assign a probability to question of the format: "chance earnings increase 12% or more"... "chance earnings increase 8%-12%"... "chance earnings decrease 0%-2%"... "chance earnings decrease 12% or higher". These questions are called 'density forecasts' and allow for a richer understanding of respondent's beliefs than point forcasts. Density forecasts are shown to elicit more complete and accurate representations about the respondent's beliefs and allows for measuring uncertaintly (Manski, 2014).

The continuous variable "Probability of decrease in earnings 12 months from now" is derived from the respondent's answer to the above probabalistic questions.

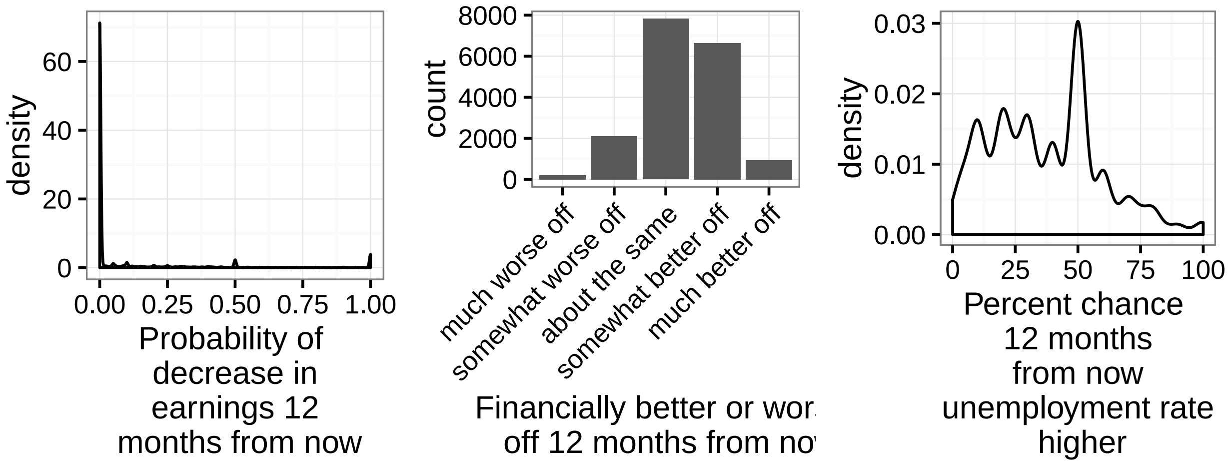
Respondents overwhelmingly stated that they expect a 0% chance that their earnings decrease in a year, only 18% of the dataset had a expectation greater than 1.

3.2.2 Response Variable 2: Financially better or worse off 12 months from now

This is a categorical response variable that results from answer to the question "Do you expect to be financially better or worse off 12 months from now?". Respondents may choose from "much better (worse) off", "somewhat better (worse) off", or "about the same".

3.2.3 Response Variable 3: Percent chance 12 months from now unemployement higher

This is a continuous response variable, generated by asking the respondent to assign a probability in response to the following question: "What is the percent chance unemployment is higher in 12 months?". Unlike response variable 1, the user can enter one answer in the form of a numeric percentage.



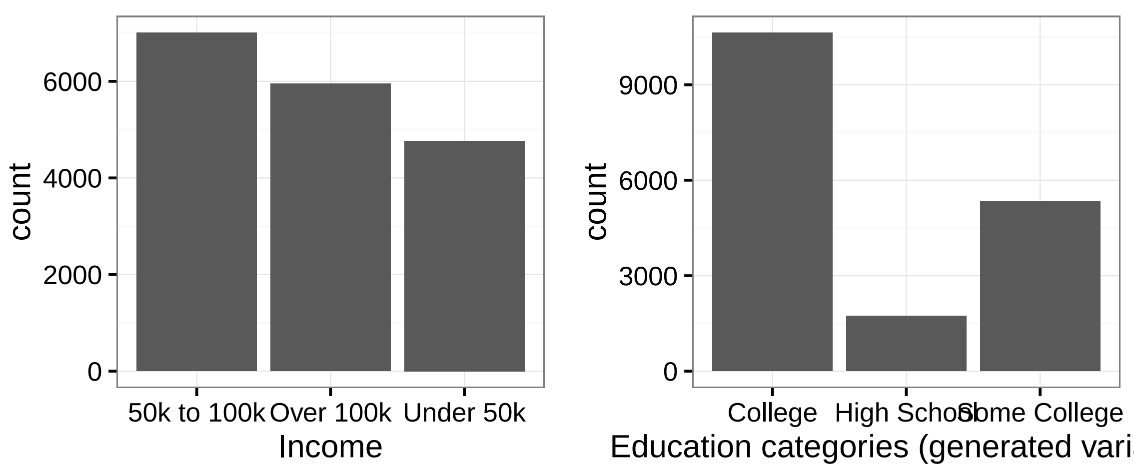
3.3 Explanatory Variables - Questions from the SCE

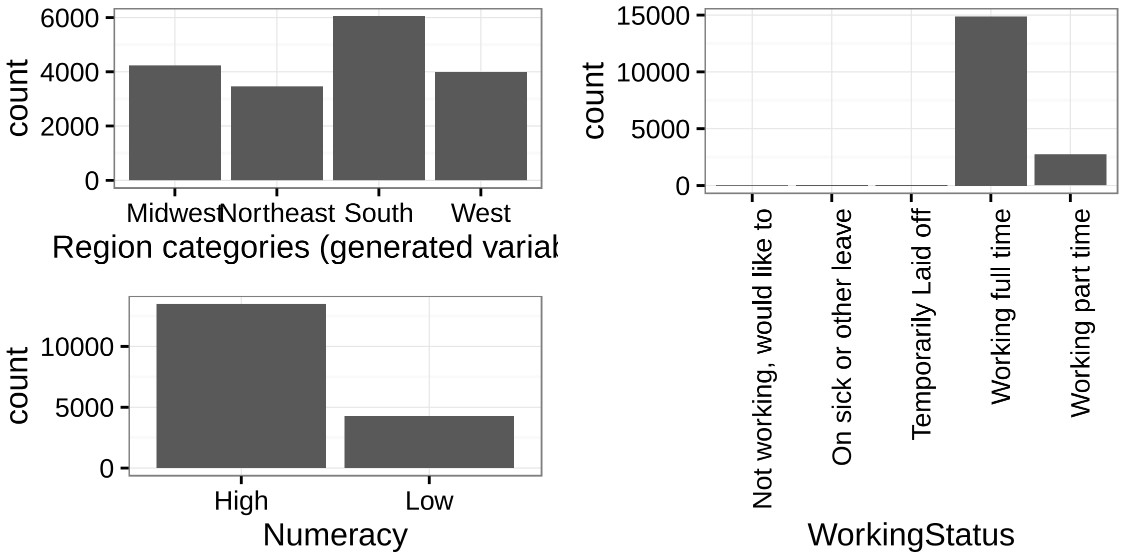
3.3.1 Age, Gender

Two explanatory variables are 'fixed', meaning they cannot be changed by the respondent - age and gender. The respondents identfied as 55% male and 45% female. The youngest respondent is 19, with a maximum age of 99. Five outliers above age 99 were removed from the dateset.

3.3.2 Income, Education, Region, Working Status, Numeracy

Several predictor variables from the SCE are changeable - income, education, region (of the U.S.), working status, and numeracy. These are all categorical (factor) variables.





3.4 Explanatory Variables - Macroeconomic Measures from FRED

The following macroeconomic measures are indicative of the status of the economy writ-large. Unlike the SCE, they are based on empirical economic measures (e.g. number of unemployment claims filed, or prices for a basket of goods) as opposed to public opinion.

Unemployement Rate (UNRATE)
The unemployment rate represents the number of unemployed as a percentage of the labor force.

Total Nonfarm Payroll (PAYEMS)
All Employees: Total Nonfarm, commonly known as Total Nonfarm Payroll, is a measure of the number of U.S. workers in the economy that excludes proprietors, private household employees, unpaid volunteers, farm employees, and the unincorporated self-employed. This measure accounts for approximately 80 percent of the population that contributes to Gross Domestic Product (GDP). This measure represent the number of jobs added or lost in an economy.

4-Week Moving Average of Initial Claims (IC4WSA)
The the 4-Week Moving Average of Initial Claims (IC4WSA) uses weekly unemployment insurance claims reported by each state's unemployment insurance program offices. These claims may be used for monitoring workload volume, assessing state program operations and for assessing labor market conditions. This 4-week moving average was aggregated to a monthly mean number of claims.

KC Fed Labor Market Conditions Index, Momentum Indicator (FRBKCLMCIM)
The Kansas City Fed Labor Market Conditions Indicator (LMCI) is a measure of labor market conditions based on 24 labor market variables.For this study we use the momentum indicator. A positive value indicates that labor market conditions are above their long-run average, while a negative value signifies that labor market conditions are below their long-run average.

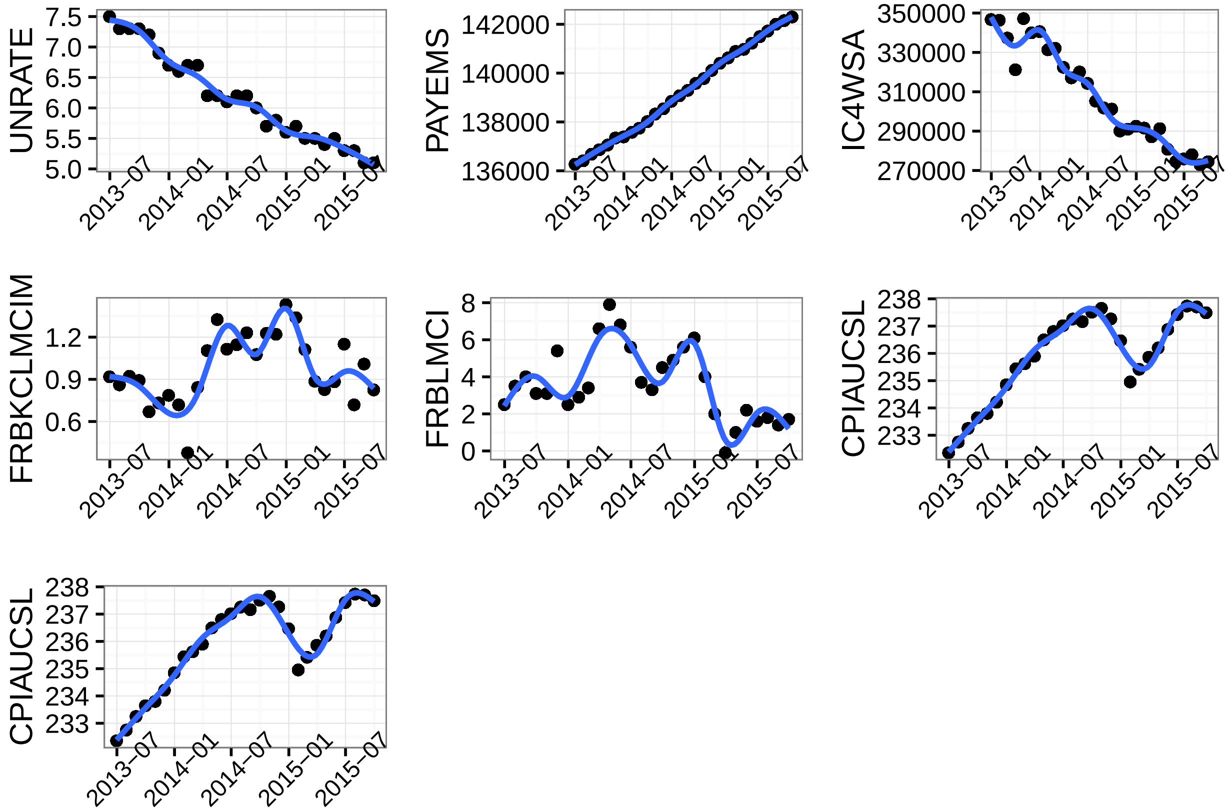
Change in Labor Market Conditions Index (FRBLMCI)

The Labor Market Conditions Index (LMCI) is derived from a dynamic factor model that extracts the primary common variation from 19, seasonally-adjusted, labor market indicators.

Consumer Price Index (CPIAUCSL)
The Consumer Price Index is a measure of the average monthly change in the price for goods and services paid by urban consumers.

S&P 500 (SP500)
The S&P500 is a gauge of the large cap U.S. equities market. The index includes 500 leading companies in

leading industries of the U.S. economy, which are publicly held on either the NYSE or NASDAQ, and covers 75% of U.S. equities.



4 Experimentation and Results

For each response variable, we build a regression model using (1) demographic variables (2) macroeconomic variables and (3) both combined demographic and macroeconomic variables. The purpose of dividing the predictor variables in such a manner was to compare whether models built using demographic variables or macroeconomic indicators would have more predictive power. Finally, a model that used variables from both groups was chosen for comparison purposes.

4.1 Response 1: Probability of decrease in earnings 12 months from now

This response variabe was transformed two way: (1) respondents who had responses implying no uncertainty (i.e. 0% or 100%) were removed and (2) a square-root transformation was performed to result in a response density more resembling a normal distribution (see Appendix I). This resulted in a sample size of approximately 4,000 responses.

Demographic Variables

This model was selected using backwards selection.

The predicted probability of a survey respondent's expected earnings decreasing was modeled using demographic variables as follows:

Probability of Decrease = .589 + .001 `Age` + .078 `High School` + .049 `Some College` - .038 `Sick or Other Leave` - .173 `Temporarily Laid Off` - .198 `Full Time` - .184 `Part Time` - .002 `Income over 100k` + .051 `Income under 50k`

Other than `Age` and the intercept, the demographic based models utilize categorical variables which are used in this model as dummy variables. For the education factors, `College` represents the baseline. For the income factors, 50 to 100k represents the baseline. Finally, for the Working Status factors, "Not working, would like to" represents the baseline.

We can take several things away from this model. Older people are very slightly more pessimistic about their earnings potential. Education other than college has a negative effect on peoples incomes. People who are in higher income brackets are more optimistic about their earning prospects. Finally, those who are either working or expecting do not expect an increase in earnings.

Macroeconomic Variables

Probability of Decrease = 1.84 - .006 `CPIAUCSL` - .045 `FRBKCLMCIM` + 0.005 `FRBLMCI`

This model was also built using backwards selection. Variance Inflation Factors were also measured, and variables were removed until all predictors had VIF values of below 10 to deal with multicollinearity.

Negative consumer price index change (`CPIAUCSL`) indicates a decreased probability of wage increase. A negative coefficient for `FRBKCLMCIM` indicates that if the labor market conditions are above their long-run average, consumers have less expectation of wage decrease. However, if that index experiences more volatility, as indicated by the positive coefficient in front of `FRBLMCI`, the predicted value of `Probability of Decrease` increases.

Combined Variables

Backward selection was used for the combined model.

Probability of Decrease = 2.04 + .001 `Age` - .007 `male` + .068 `NumeracyLow` - .003 `Northeast` + .008 `South` + .045 `West` + .050 `High School` + .035 `Some College` + .005 `Income over 100k` + .042 `Income under 50k` - .007 `CPIAUCSL` - .048 `FRBKCLMCIM` + 0.006 `FRBLMCI`

The regional factor (a demographic variable) was included for signficance in this model, as opposed to the pure demographic model that did not include the factor. The baseline for this factor was `Midwest`. The `Northeast` had a negative coefficient, while the `South` and `West` had positive coefficients, indicating less pessimism in the Northeast.

Also, gender was included in this model. The baseline was `female`. `Male` had a negative coefficient. Finally, `Numeracy` was included as a factor, where the baseline was `High`. `Low` also had a positive coefficient, indicating consumers with low numeracy scores were more likely to be pessimistic.

None of the `Working Status` factors were found to be signficant in the combined model, thus that factor was not included.

Beyond the additional introduced predictors, all other predictors had the same sign on their coefficient as the previous two models, with one notable exception: in the combined model, the effect of `Income over 100k` reversed as the coefficients sign changed from negative to positive. Thus the combined model associates making more income with more pessimism.

4.2 Response 2: Financially better or worse off 12 months from now

For the response variable "Financially better or worse off 12 months from now" a multinomial logit model was fit to the factor response. Stepwise backwards selection was used to remove insignficant variables until a model with all significant variables was found.

Demographic Variables
The result of using stepwise backward selection on demographic variables resulted in a model using: Current age, Gender, Numeracy, Region categories, Education categories, and Income. McFadden's pseudo R2 for this model is 0.030, indicating a fairly poor fit.

Demographic models using on significant coefficients:
(1)"much worse off" $= -4.41 + 0.051 currentage - 0.29 numeracy + 0.37 education : highschool + 1.07 income :$

$under50k$
(2)"somewhat worse off" $= -2.47 + 0.073 currentage - 0.22 numeracy - 0.46 region : south - 0.36 region : west + 0.20 income : over100k + 0.40 income : under50k$
(3)"about the same" $= 0.14 + 0.047 currentage - 0.16 numeracy - 0.43 region : south - 0.24 region : west + 0.31 income : under50k$
(4)"somewhat better off" $= 1.14 + 0.02 currentage + 0.24 gender - male - 0.32 numeracy - 0.30 region : northeast - 0.41 region : south - 0.42 region : west - 0.21 education : highschool + 0.01 income : over100k + 0.33 income : under50k$
(5)"much better off" = 1 - "much worse off" - "somewhat worse off" - "about the same" - "somewhat better off"

Macroeconomic Variables
The result of using stepwise backward selection on macroeconomic variables resulted in a model using: only the unemployment rate variable. McFadden's pseudo R2 for this model is 0.003, indicating a fairly poor fit.
Macroeconomic models using on significant coefficients:
(1)"much worse off" $= -6.38 + 0.79 unemployement.rate$
(2)"somewhat worse off" $= -1.93 + 0.44 unemployment.rate$
(3)"about the same" $= 0.43 + 0.27 unempoyment.rate$
(4)"somewhat better off" $= 1.01 + 0.14 unemployment.rate$
(5)"much better off" = 1 - "much worse off" - "somewhat worse off" - "about the same" - "somewhat better off"

Combined Variables
The result of using stepwise backward selection when combinging all available variables. McFadden's pseudo R2 for this model is 0.034, indicating a fairly poor fit but slightly better than the demographic variables alone.

Combined models using on significant coefficients:
(1)"much worse off" $= -9.60 + 0.05 currentage + 0.38 education : highschool + 1.07 income : under50k + 0.31 unemployment$
(2)"somewhat worse off" $= -5.37 + 0.073 currentage - 0.22 numeracy - 0.29 region : northeast - 0.46 region : south - 0.36 region : west + 0.18 education : highschool + 0.20 income : over100k + 0.40 income : under50k + 0.47 unemployment$
(3)"about the same" $= -1.60 + 0.047 currentage - 0.16 numeracy - 0.45 region : south - 0.25 region : west + 0.30 income : under50k + 0.28 unemployment$
(4)"somewhat better off" $= 0.23 + 0.02 currentage + 0.24 gender - male - 0.32 numeracy - 0.30 region : northeast - 0.41 region : south - 0.43 region : west - 0.21 education : highschool + 0.22 income : over100k + 0.33 income : under50k + 0.15 unemployment$
(5)"much better off" = 1 - "much worse off" - "somewhat worse off" - "about the same" - "somewhat better off"

Age, gender, numeracy, income, and education rate had almost identical coefficients as the demographic model. Region, however, had different coefficient values and different significance for a given response variable level. Unemployment was much lower (but still signficant) in the combined model, owing to the influence of sigificant variables in the combined model, as opposed to the macroeconomic variable model when it was the only variable used.

4.3 Response 3: Percent chance 12 months from now unemployment higher

For responses to the question of "Percent chance 12 months from now unemployment higher", we fit a linear model including all predictor variables, only demographic variables, and only macroeconomic variables with the response variable.

Demographic Variables

`Percent chance 12 months from now Unemployment Rate Higher` = 35.056 + 3.4967 `Male` + 2.297 `High School` + 2.042 `Some College` - 2.002 `IncomeOver 100k` + .696 `IncomeUnder 50k`

This model was chosen using backward selection, and the baseline for dummy variables is the same as the demographic variable models. Being male, non college educated, and in a lower income bucket predicts the respondent will be more likely to predict that unemployment rates will rise. It's possible that the more likely the respondent is to be unemployed, the more likely they thing unemployment rates will rise. This assertion, however would need to be verified by further research.

Macroeconomic Variables

`Percent chance 12 months from now Unemployment Rate Higher` = -103.55 + 0.758 `CPIAUCSL` - 3.970 `FRBKCLMCIM` - 0.018 `SP500`

The model using macro indicators was also built using backward selection. An increase in the consumer price index (`CPIAUCSL`) predicts a raise in expected unemployment rates, while decrease in labor market employement levels `FRBKCLMCIM` and the market `SP500` has the opposite effect. Of note is that `UNRATE` was not a signficant predictor of consumers unemployment rate expectations.

Combined Variables

`Percent chance 12 months from now Unemployment Rate Higher` = -357.4 + 3.134 `male` -2.83 `NumeracyLow` + .076 `Northeast` + 1.328 `South` + .440 `West` + 3.11 `High School` + 2.47 `Some College` - 2.247 `IncomeOver 100k` + 1.082 `Income Under 50k` + 1.049 `CPIAUCSL` - 3.65 `FRBKCLMCIM` + .001 `PAYEMS` - .013 `SP500` .00003 + `IC4WSA` + 3.324 `UNRATE`

This model was chosen via backward selection.

Perhaps the most notable change from the combined model from either the demographic or macro model is that `UNRATE` is not only included, but also exhibits the strongest relationship. `IC4WSA` is also introduced, albeit with a very small coefficient. From the demographic set, Numeracy and Region are also found to have a signficant impact thus are added.

4.4 Summary of Responses

Below is a table showing various methods of model evaluation for each model built above. "r[1-3]" refers to the reponse value being measured. "Demo", "Macro", and "Comb" refer to the predictor variable set, either demographic, macroeconomic, or a combination of the two, respectively.

Model	SD	SE	AIC	BIC	LogLik
r1_demo_final	0.38	0.44	-455.48	-385.76	238.74
r1_macro_final	0.38	0.44	-301.28	-269.59	155.64
r1_comb_final	0.38	0.43	-541.64	-446.57	285.82
r2_demo_final	NA	NA	38350.25	38692.67	-19131.12
r2_macro_final	NA	NA	39337.10	39399.36	-19660.55
r2_comb_final	NA	NA	38211.99	38585.53	-19057.99
r3_demo_final	1.90	37.27	160069.44	160123.92	-80027.72
r3_macro_final	2.00	37.20	160096.97	160135.89	-80043.49
r3_comb_final	2.84	37.25	159822.72	159955.02	-79894.36

5 Discussion and Conclusions

For each response variable, we find that that demographic variables are better predictors of consumer sentiment than macroeconomic indicators. This reinforces the finding of Kuchler and Zafar that consumers opinion of the economy has more to do with elements that are visible to the consumer, than overall aggregate indicators. Future research could compare whether the demographic variables used in this study perform better than the local housing market. Beyond economics, it may warrant further exploration of how an individuals situation can override aggregate mesaures in other settings.

We also find that the the response variable `Probability of decrease in earnings 12 months from now` has the ability to be modeled with less information loss than either of the the other 2 response variables (`Financially better or worse off 12 motnhs from now` or `Percent chance 12 motnhs from now unemployment rate higher`). Again, this indicates individuals have a clearer view of their own personal situation than of the economic climate as a whole.

Finally, we can broadly summarize the effects of the different variables from both sets on consumer outlook. From the demographic, we find males, those that do not hold a college degree, those with lower incomes, and those not from the Northeast tend to be generally more pessimistic about future economic outcomes to varying degrees.

For the macroeconomic indicators, higher prices and higher labor market variance are associated with more pessimism, while increases in labor market rates and the SP500 tend to be associated with less pessimism. However, these effects are not as strong as those observed by the demogrpahic set.

6 References

Board of Governors of the Federal Reserve System (US), Change in Labor Market Conditions Index [FRBLMCI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/FRBLMCI, July 16, 2016.

Federal Reserve Bank of Kansas City, KC Fed Labor Market Conditions Index, Momentum Indicator [FRBKCLMCIM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/FRBKCLMCIM, July 15, 2016.

Ludvigson, Sydney C.(2004). Consumer Confidence and Consumer Spending. Journal of Economic Perspectives 18(2). 29-50. doi:10.1257/0895330041371222

Manski, C. F. (2014). Measuring Expectations. Econometrica, 72(5), 1329-1376. doi:10.1111/j.1468-0262.2004.00537.x

S&P Dow Jones Indices LLC, S&P 500© [SP500], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, July 15, 2016.

US. Employment and Training Administration, 4-Week Moving Average of Initial Claims [IC4WSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/IC4WSA, July 16, 2016.

US. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, July 16, 2016.

US. Bureau of Labor Statistics, All Employees: Total Nonfarm Payrolls [PAYEMS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PAYEMS, July 16, 2016.

US. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPIAUCSL, July 16, 2016.

US. Bureau of Labor Statistics, Other Separations: Total Nonfarm [JTSOSL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/JTSOSL, July 15, 2016.

Doms, M., & Marin, N. (2004, July). Consumer Sentiment, the Economy, and the News Media. Retrieved July 10, 2016, from Www.baruch.cuny.edu/library/index.php.

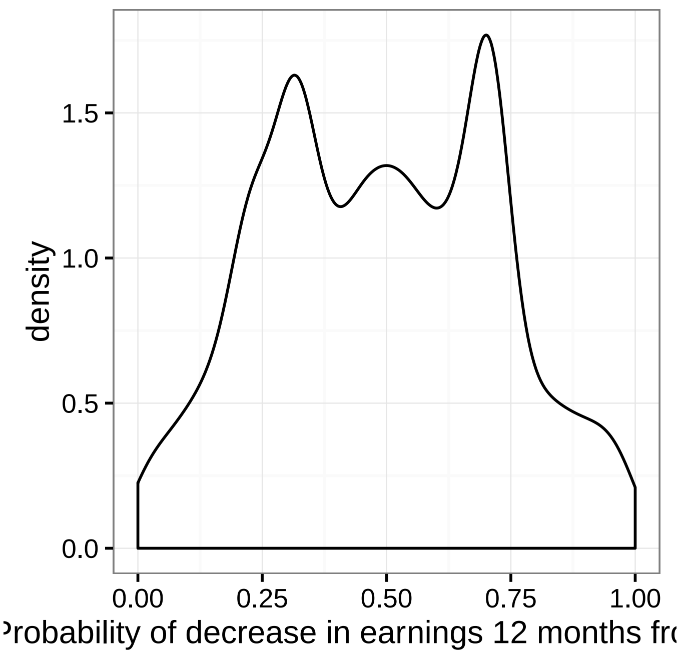
Kuchler, T., & Zafar, B. (2015, October). Personal Experiences and Expectations about Aggregate Outcomes. FRBNY Staff Reports, 748. Retrieved July 10, 2016, from Www.baruch.cuny.edu/library/index.php.

Spulveda-Umanzor, J. (2004, April). The Relation Between Macroeconomic Uncertainty And The Expected Performance of the Economy. Retrieved July 10, 2016, from Www.baruch.cuny.edu/library/index.php.

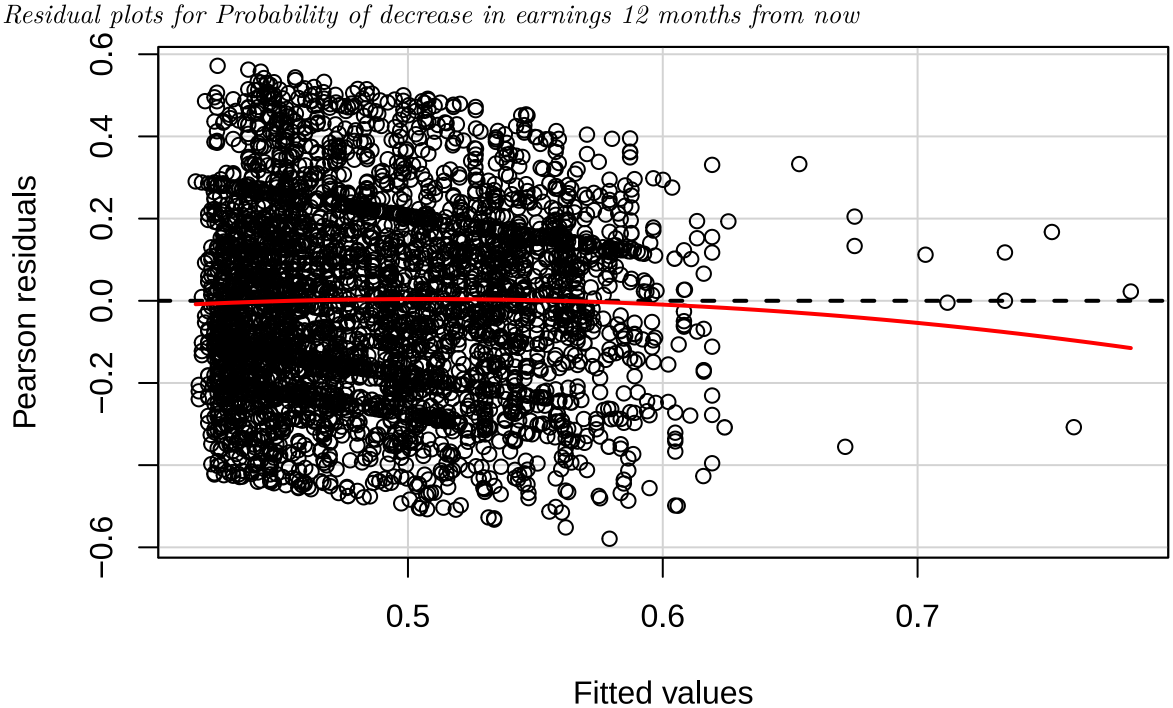
Souleles, N. S. (2004, February). Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence form the Michigan Consumer Sentiment Surveys. Journal of Money, Credit, and Banking, 36(1), 39-72. Retrieved July 10, 2016, from Www.baruch.cuny.edu/library/index.php.

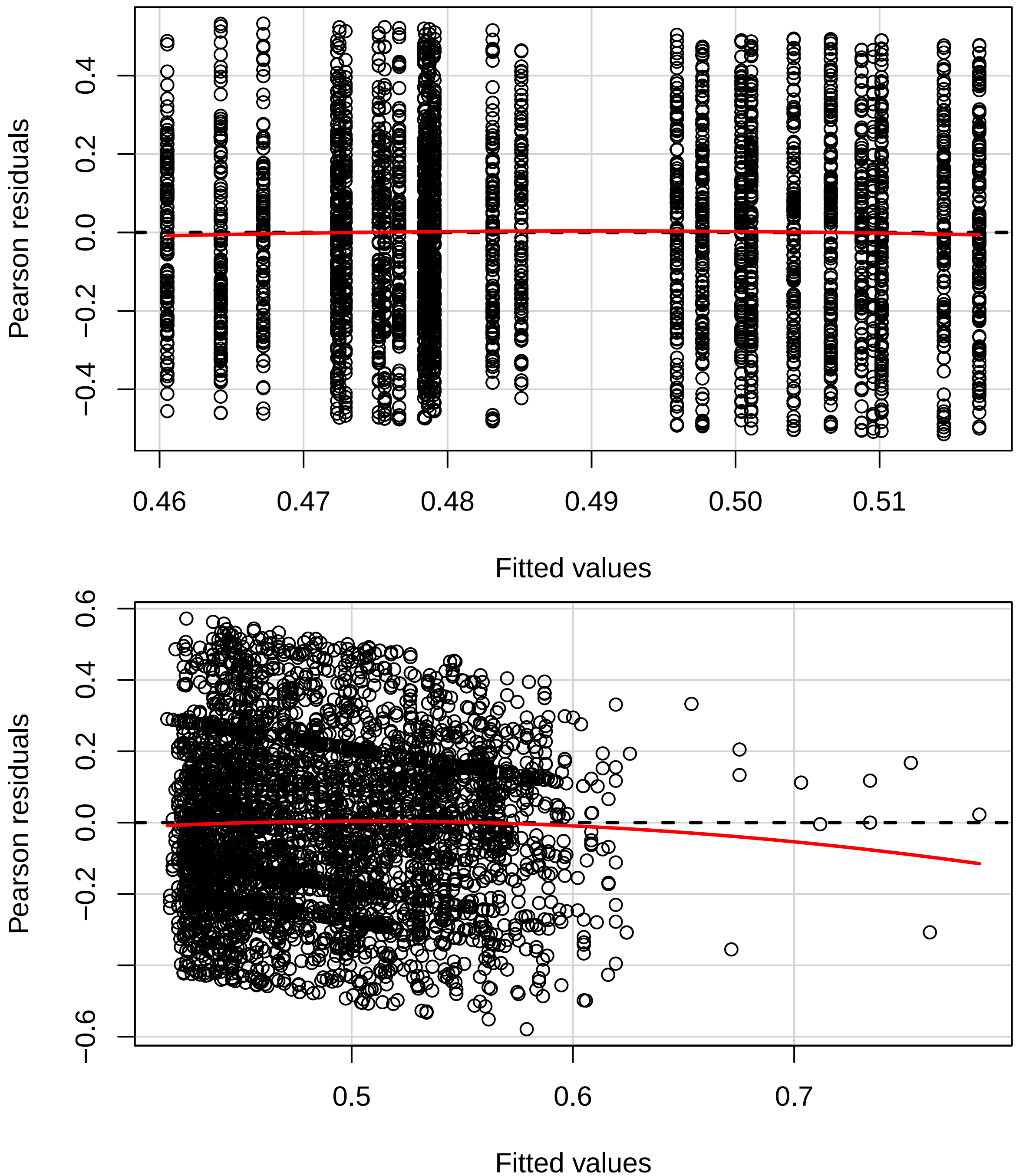
7 Appendix I - Supplemental Tables and Figures

Density plot of adjusted response variable 1: `Probability of decrease in earnings 12 months from now`

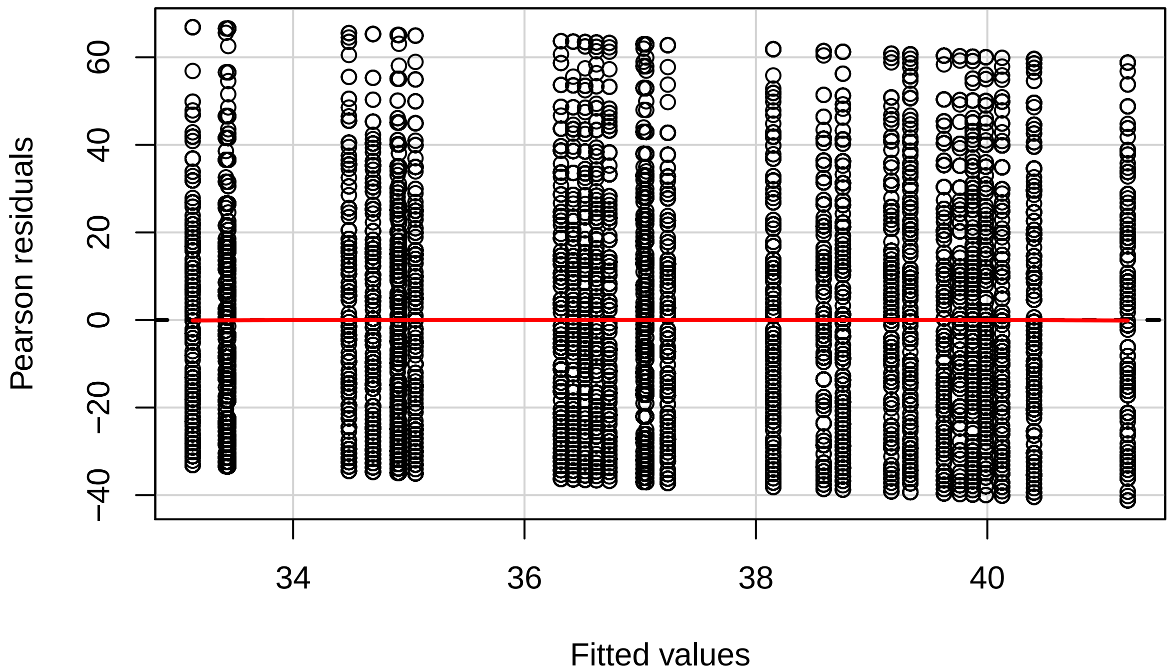


Residual plots for Probability of decrease in earnings 12 months from now





Residual plots for Percent chance 12 months from now unemployment rate higher



8 Appendix II - R statistical programming code

```
library(dplyr)
library(ggplot2)
library(knitr)
library(VGAM)
library(car)

require(grid)
require(gridExtra)
final.data.all <- readRDS("final_data.Rda")

#levels(x)[levels(x)=="beta"] <- "two"

levels(final.data.all$Gender)[levels(final.data.all$Gender) == ""] <- NA
levels(final.data.all$Numeracy)[levels(final.data.all$Numeracy) == ""] <- NA
levels(final.data.all$`Region categories (generated variable)`)[levels(final.data.all$`Region categories
levels(final.data.all$`Education categories (generated variable)`)[levels(final.data.all$`Education cate
levels(final.data.all$WorkingStatus)[levels(final.data.all$WorkingStatus) == ""] <- NA
levels(final.data.all$Income)[levels(final.data.all$Income) == ""] <- NA


final.data <- final.data.all
# here, select the variables we're using for the analysis
# all should have a section in the explanations below
final.data.all <- na.omit(final.data.all)
final.data <- na.omit(final.data)

predictor.variables = c("Probability of decrease in earnings 12 months from now",
```

```
                        "Financially better or worse off 12 months from now",
                        "Percent chance 12 months from now unemployment rate higher",
                        "Current age",
                        "Gender",
                        "Numeracy",
                        "Region categories (generated variable)",
                        "Education categories (generated variable)",
                        "WorkingStatus",
                        "Income",
                        "CPIAUCSL",
                        "FRBKCLMCIM",
                        "FRBLMCI",
                        "PAYEMS",
                        "SP500",
                        "IC4WSA",
                        "UNRATE")

final.data <- dplyr::select(final.data, one_of(predictor.variables))

#final.data$`Financially better or worse off 12 months from now` <- as.ordered(final.data$`Financially

set.seed(5648)

which.train <- sample(x = c(TRUE, FALSE), size = nrow(final.data), replace = TRUE, prob = c(0.8, 0.2))

final.data <- final.data[which.train, ]

final.data.all <- final.data.all[which.train, ]


final.data.test <- final.data[!which.train, ]
final.data.test <- na.omit(final.data.test)
final.data.test <- as.data.frame(final.data.test)
total.responses <- length(final.data.all$`Probability of decrease in earnings 12 months from now`)
prob.decrease.non.zero.responses <- filter(final.data.all, `Probability of decrease in earnings 12 month

non.zero.responses <- length(prob.decrease.non.zero.responses$`Probability of decrease in earnings 12 mo

proportion.non.zero <- non.zero.responses / total.responses

# show density plot
prob.decrease.1 <- ggplot(final.data, aes(x =`Probability of decrease in earnings 12 months from now`))
prob.betterworse.1 <- ggplot(final.data, aes(x = `Financially better or worse off 12 months from now`))
prob.unemp.1 <- ggplot(final.data, aes(x =`Percent chance 12 months from now unemployment rate higher`))
grid.arrange(prob.decrease.1, prob.betterworse.1, prob.unemp.1, ncol = 3, nrow = 1)
male.prop <- (final.data %>% filter(Gender == "male") %>% count()) / length(final.data$Gender)
female.prop <- (final.data %>% filter(Gender == "female") %>% count()) / length(final.data$Gender)


final.data$`Current age`[final.data$`Current age` > 100]  <- NA
a <- summary(final.data$`Current age`)
sce.variable.set <- c("Income", "Education categories (generated variable)", "Region categories (generat
#kable(summary(dplyr::select(final.data, one_of(sce.variable.set ))))
```

```
income <- ggplot(final.data, aes(x = Income)) + geom_bar() + theme_bw()
edu <- ggplot(final.data, aes(x = `Education categories (generated variable)`)) + geom_bar() + theme_bw
region <- ggplot(final.data, aes(x = `Region categories (generated variable)`)) + geom_bar() + theme_bw
working <- ggplot(final.data, aes(x = WorkingStatus)) + geom_bar() + theme_bw() + theme(axis.text.x = el
numeracy <- ggplot(final.data, aes(x = Numeracy)) + geom_bar() + theme_bw()

grid.arrange(income, edu, ncol = 2, nrow = 1)
grid.arrange(grid.arrange(region, numeracy, ncol = 1, nrow = 2), working, ncol = 2, nrow = 1)

a <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `UNRATE`)) + geom_point() + ge
b <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `PAYEMS`)) + geom_point() + ge
c <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `IC4WSA`)) + geom_point() + ge
d <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `FRBKCLMCIM`)) + geom_point() +
e <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `FRBLMCI`)) + geom_point() + g
f <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `CPIAUCSL`)) + geom_point() +
g <- ggplot(final.data.all, aes(x = `Month Survey was administered`, y = `CPIAUCSL`)) + geom_point() +

grid.arrange(a, b, c, d, e, f, g, ncol = 3, nrow = 3)
#splitting the data into separate data frames for each model
#response variable 1 (Probability of decrease in earnings 12 months from now)
final.data$`Probability of decrease in earnings 12 months from now` <- sqrt(final.data$`Probability of d
r1macro_data <- final.data[,c(1,11:17)]

r1macro_data <- r1macro_data %>% filter(`Probability of decrease in earnings 12 months from now` != 0) %


r1demo_data <- final.data[,c(1,4:10)]

r1demo_data  <- r1demo_data  %>% filter(`Probability of decrease in earnings 12 months from now` != 0) %

r1comb_data <- final.data[,c(1,4:17)]

r1comb_data <- r1comb_data  %>% filter(`Probability of decrease in earnings 12 months from now` != 0) %>

#response 3 (Percent chance 12 motnhs from now unemployment will be higher)
r3macro_data <- final.data[,c(3,11:17)]
r3demo_data <- final.data[,c(3,4:10)]
r3comb_data <- final.data[,c(3,4:17)]
#Response 1: Demo predictors
r1_demo_no_region_no_g <- lm(`Probability of decrease in earnings 12 months from now` ~ . -`Region categ
#summary(r1_demo_no_region_no_g)

r1_demo_final <- r1_demo_no_region_no_g
#Reponse 1: Macro predictors
r1_macro_m4 <- lm(`Probability of decrease in earnings 12 months from now` ~ . -`IC4WSA` -`SP500` -`UNRA
#summary(r1_macro_m4)
#vif(r1_macro_m4)

r1_macro_final <- r1_macro_m4
#summary(r1_macro_final)
#Response 1: combined predictors
r1_comb_final <- lm(`Probability of decrease in earnings 12 months from now` ~ . -`UNRATE` -`PAYEMS` -`I
#summary(r1_comb_final)
```

```r
remove.these <- c("Probability of decrease in earnings 12 months from now", "Percent chance 12 months fr
analysis.data.2 <- final.data %>% dplyr::select(-one_of(remove.these))

r2_demo_final <- vglm(`Financially better or worse off 12 months from now`~ `Current age` + Gender + Num

demo.2.null <- vglm(`Financially better or worse off 12 months from now`~ 1, family = multinomial, data

LLf   <- VGAM::logLik(r2_demo_final)
LL0   <- VGAM::logLik(demo.2.null )
mcf.demo <- as.vector(1 - (LLf / LL0))
r2_macro_final <- vglm(`Financially better or worse off 12 months from now`~ UNRATE, family = multinomia

macro.2.null <- vglm(`Financially better or worse off 12 months from now`~ 1, family = multinomial, data

LLf   <- VGAM::logLik(r2_macro_final)
LL0   <- VGAM::logLik(macro.2.null)
mcf.macro <- as.vector(1 - (LLf / LL0))
r2_comb_final <- vglm(`Financially better or worse off 12 months from now`~ `Current age` + Gender + Num

comb.2.null <- vglm(`Financially better or worse off 12 months from now`~ 1, family = multinomial, data

LLf   <- VGAM::logLik(r2_comb_final )
LL0   <- VGAM::logLik(comb.2.null)
mcf.comb <- as.vector(1 - (LLf / LL0))
#Response 3: Demo predictors

r3_demo_no_region_no_age <- lm(`Percent chance 12 months from now unemployment rate higher` ~ . -`Regio

#summary(r3_demo_no_region_no_age)

r3_demo_final <- r3_demo_no_region_no_age
#Response 3: Macro predictors
r3_macro_final <- lm(`Percent chance 12 months from now unemployment rate higher` ~ . -`FRBLMCI` -`UNRA

#summary(r3_macro_final)
#vif(r3_macro_final)

r3_comb_final <- lm(`Percent chance 12 months from now unemployment rate higher` ~ . -`WorkingStatus` -

#summary(r3_comb_final)
calc_sd <- function(fit, data, target){
  prediction <- predict(fit, newdata=data, type='response')
  difference <- (prediction - mean(data[,target]))
  difference_squared <- difference * difference
  return (mean(sqrt(difference_squared)))
}

calc_se <- function(fit, data, target){
  prediction <- predict(fit, newdata=data, type='response')
  difference <- (prediction - data[,1])
  difference_squared <- difference * difference
  return (mean(sqrt(difference_squared)))
}
```

```r
# Put in readable model names at the end!!!
Model <- c("r1_demo_final",
           "r1_macro_final",
           "r1_comb_final",
           "r2_demo_final",
           "r2_macro_final",
           "r2_comb_final",
           "r3_demo_final",
           "r3_macro_final",
           "r3_comb_final")

SD <- format(c(calc_sd(r1_demo_final,final.data.test, 1),
               calc_sd(r1_macro_final,final.data.test, 1),
               calc_sd(r1_comb_final,final.data.test, 1), NA, NA, NA,
               #calc_sd(r2_demo_final,final.data.test, 2),
               #calc_sd(r2_macro_final,final.data.test, 2),
               #calc_sd(r2_comb_final,final.data.test, 2),
               calc_sd(r3_demo_final,final.data.test, 3),
               calc_sd(r3_macro_final,final.data.test, 3),
               calc_sd(r3_comb_final,final.data.test, 3)
               ), digits=2, nsmall=2)

SE <- format(c(calc_se(r1_demo_final,final.data.test, 1),
               calc_se(r1_macro_final,final.data.test, 1),
               calc_se(r1_comb_final,final.data.test, 1), NA, NA, NA,
               #calc_se(r2_demo_final,final.data.test, 2),
               #calc_se(r2_macro_final,final.data.test, 2),
               #calc_se(r2_comb_final,final.data.test, 2),
               calc_se(r3_demo_final,final.data.test, 3),
               calc_se(r3_macro_final,final.data.test, 3),
               calc_se(r3_comb_final,final.data.test, 3)
               ), digits=2, nsmall=2)
AIC <- format(c(AIC(r1_demo_final),
                AIC(r1_macro_final),
                AIC(r1_comb_final),
                AIC(r2_demo_final),
                AIC(r2_macro_final),
                AIC(r2_comb_final),
                AIC(r3_demo_final),
                AIC(r3_macro_final),
                AIC(r3_comb_final)
                ), digits=2, nsmall=2)

BIC <- format(c(BIC(r1_demo_final),
                BIC(r1_macro_final),
                BIC(r1_comb_final),
                BIC(r2_demo_final),
                BIC(r2_macro_final),
                BIC(r2_comb_final),
                BIC(r3_demo_final),
                BIC(r3_macro_final),
                BIC(r3_comb_final)
                ), digits=2, nsmall=2)
```

```
LogLik <- format(c(logLik(r1_demo_final),
                   logLik(r1_macro_final),
                   logLik(r1_comb_final),
                   logLik(r2_demo_final),
                   logLik(r2_macro_final),
                   logLik(r2_comb_final),
                   logLik(r3_demo_final),
                   logLik(r3_macro_final),
                   logLik(r3_comb_final)
                   ), digits=2, nsmall=2)

kable(cbind(Model, SD, SE, AIC, BIC, LogLik))
ggplot(r1macro_data, aes(x = `Probability of decrease in earnings 12 months from now`)) + geom_density(
residualPlot(r1_demo_final)
residualPlot(r1_macro_final)
residualPlot(r1_demo_final)
#residualPlot(r3_demo_final) #results in differing lengths of resid/values
residualPlot(r3_macro_final)
#residualPlot(r3_comb_final)
##
```