MATH 212 Project Preliminary Analysis

Zack Roder

4/18/2021

Data Set

From https://fred.stlouisfed.org/, I retrieved a data set with various important economic indicators over the period from Q2 2011 through Q1 2020 (immediately before the COVID-induced recession).

```
econ_indicators
```

```
## # A tibble: 36 x 6
##    observation_date    UNRATE    GDP MEDIAN_WEELKY_WAGES  DJCA HOUST
##    <dttm>               <dbl>  <dbl>               <dbl> <dbl> <dbl>
##  1 2011-04-01 00:00:00    9.1 15496.                 754 4271.  1723
##  2 2011-07-01 00:00:00    9   15592.                 760 4008.  1858
##  3 2011-10-01 00:00:00    8.5 15796.                 760 4068.  2015
##  4 2012-01-01 00:00:00    8.2 16020.                 764 4377.  2122
##  5 2012-04-01 00:00:00    8.2 16152.                 772 4365.  2218
##  6 2012-07-01 00:00:00    7.8 16257.                 766 4435.  2341
##  7 2012-10-01 00:00:00    7.9 16359.                 771 4413.  2724
##  8 2013-01-01 00:00:00    7.5 16570.                 768 4792.  2860
##  9 2013-04-01 00:00:00    7.5 16638.                 777 5091.  2604
## 10 2013-07-01 00:00:00    7.2 16849.                 779 5181.  2647
## # ... with 26 more rows
```

1. There are 5 variables in my data set.

- Unemployment rate (UNRATE) - From FRED: "The unemployment rate represents the number of unemployed as a percentage of the labor force. Labor force data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces. This rate is also defined as the U-3 measure of labor underutilization." Unit is percent (seasonally adjusted).
- Gross domestic product (GDP) - From FRED: "Gross domestic product (GDP), the featured measure of U.S. output, is the market value of the goods and services produced by labor and property located in the United States." Unit is billions of dollars, seasonally adjusted annual rate.
- Employed full time: median usual weekly nominal earnings, 16 years and over (MEDIAN_EARNINGS) - from FRED: "Data measure usual weekly earnings of wage and salary workers. Wage and salary workers are workers who receive wages, salaries, commissions, tips, payment in kind, or piece rates. The group includes employees in both the private and public sectors but, for the purposes of the earnings series, it excludes all self-employed persons, both those with incorporated businesses and those with unincorporated businesses. Usual weekly earnings represent earnings before taxes and other deductions and include any overtime pay, commissions, or tips usually received (at the main job in the case of multiple jobholders). Prior to 1994, respondents were asked how much they usually earned per week. Since January 1994, respondents have been asked to identify the easiest way for them to report earnings (hourly, weekly, biweekly, twice monthly, monthly, annually, or other) and how much they usually earn in the reported time period. Earnings reported on a basis other than weekly are converted to a weekly

equivalent. The term"usual" is determined by each respondent's own understanding of the term. If the respondent asks for a definition of "usual," interviewers are instructed to define the term as more than half the weeks worked during the past 4 or 5 months." Unit is dollars, seasonally adjusted.

- Dow Jones Composite Average (DJCA) - From FRED: "The Dow Jones Composite Average is combination of all three major Dow Jones Averages (Industrial, Utility, and Transportation). Since the Composite Average is made up of this select group of prominent stocks, Dow Jones refers to it as a blue chip microcosm of the US stock market." Unit is index, not seasonally adjusted.
- Housing Starts (HOUST) - From FRED: "As provided by the Census, start occurs when excavation begins for the footings or foundation of a building. All housing units in a multifamily building are defined as being started when this excavation begins. Beginning with data for September 1992, estimates of housing starts include units in structures being totally rebuilt on an existing foundation."

2. My response variable in median weekly wages, while my predictor variables are unemployment, GDP, DJCA, and housing starts.

3. My research question is: can we linearly predict median weekly wages given some combination of unemployment, GDP, DJCA, and housing starts? I think this is an interesting question because median weekly wages provides a stronger indication of the economic conditions for actual working people in America. GDP, while important, doesn't tell us precisely if people are thriving or not. Hence, I want to see how median wages can be predicted from these other important economic indicators.

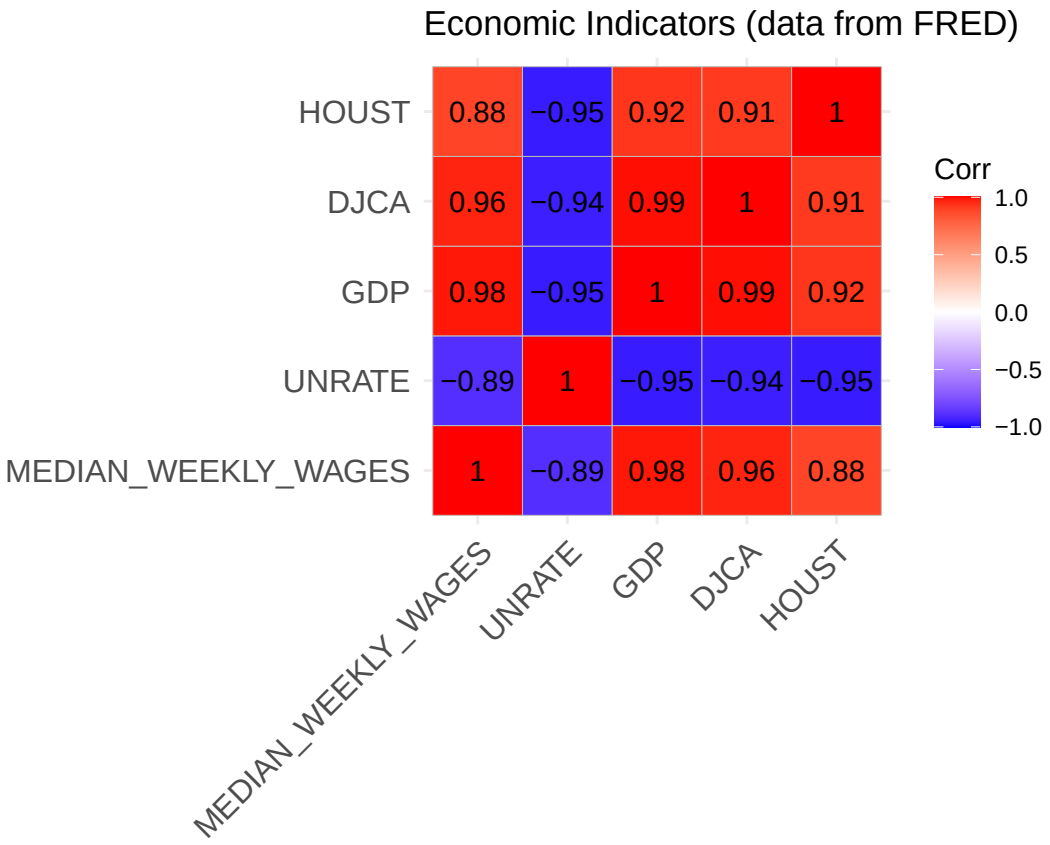
Correlation Analysis

1. Create correlation plot.

```
MEDIAN_WEEKLY_WAGES <- econ_indicators$MEDIAN_WEELKY_WAGES
UNRATE <- econ_indicators$UNRATE
GDP <- econ_indicators$GDP
DJCA <- econ_indicators$DJCA
HOUST <- econ_indicators$HOUST

DF_indicators <- data.frame(MEDIAN_WEEKLY_WAGES, UNRATE, GDP, DJCA, HOUST)
r <- cor(DF_indicators)

ggcorrplot(r, title="Economic Indicators (data from FRED)", lab=TRUE)
```



2. The matrix indicates very high correlation between MEDIAN_WEEKLY_WAGES and both GDP and DJCA (0.98 and 0.96). Either of these predictor variables seem to be good candidates for being the first to enter our stepwise model.
3. All of our predictor values are correlated with our response variable (MEDIAN_WEEKLY_WAGES). UNRATE is negatively correlated, while HOUST, GDP, and DJCA are all positively correlated with our response variable.
4. Given how closely related some of these indicators are, we, unsurprisingly, do see multicolinearity in the dataset. For example, we see that GDP and DJCA are highly correlated (0.99).

Variable Selection

1. Regress MEDIAN_WEEKLY_WAGES on each predictor variable and compare outputs. Use alpha_to_enter = 0.15 and alpha_to_remove = 0.15.

```
model1 <- lm(MEDIAN_WEEKLY_WAGES ~ DJCA, data=DF_indicators)
summary(model1)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ DJCA, data = DF_indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.148  -8.956   5.366   9.865  41.847
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.068e+02  1.083e+01   56.04   <2e-16 ***
## DJCA        3.433e-02  1.639e-03   20.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.41 on 34 degrees of freedom
## Multiple R-squared:  0.9281, Adjusted R-squared:  0.926
## F-statistic:   439 on 1 and 34 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(MEDIAN_WEEKLY_WAGES ~ GDP, data=DF_indicators)
summary(model2)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP, data = DF_indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.242  -5.412  -1.048   6.200  32.190
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.930e+02  1.731e+01   16.92   <2e-16 ***
## GDP         2.893e-02  9.327e-04   31.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.62 on 34 degrees of freedom
## Multiple R-squared:  0.9659, Adjusted R-squared:  0.9649
## F-statistic: 962.4 on 1 and 34 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(MEDIAN_WEEKLY_WAGES ~ UNRATE, data=DF_indicators)
summary(model3)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ UNRATE, data = DF_indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.204 -19.608  -3.611  15.235  85.521
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  990.160     14.999   66.01  < 2e-16 ***
## UNRATE       -28.791      2.536  -11.36 4.09e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.25 on 34 degrees of freedom
## Multiple R-squared:  0.7913, Adjusted R-squared:  0.7852
## F-statistic: 128.9 on 1 and 34 DF,  p-value: 4.087e-13
```

```
model4 <- lm(MEDIAN_WEEKLY_WAGES ~ HOUST, data=DF_indicators)
summary(model4)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ HOUST, data = DF_indicators)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.95 -21.14  -6.13  21.36  45.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.906e+02  2.185e+01   27.03  < 2e-16 ***
## HOUST       7.409e-02  6.695e-03   11.07 8.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.79 on 34 degrees of freedom
## Multiple R-squared:  0.7827, Adjusted R-squared:  0.7763
## F-statistic: 122.5 on 1 and 34 DF,  p-value: 8.189e-13
```

Since the p-value for each of the predictor variables' slopes is less than our alpha_to_enter, each is a candidate to be entered into the stepwise model. We choose the one with the lowest p-value. Both the MEDIAN_WEEKLY_WAGES ~ GDP and MEDIAN_WEEKLY_WAGES ~ DJCA have slopes with p-value < 2e-16, but we notice that the t-value for the slope in the GDP model is higher, indicating greater significance. Thus, we'll choose GDP as the first variable for our model.

2. Fit each of the two-predictor models that include GDP as a predictor.

```
model1 <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + DJCA, data=DF_indicators)
summary(model1)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + DJCA, data = DF_indicators)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.9655  -6.9021  -0.1618   6.3199  30.0933
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 201.530534  63.562538   3.171  0.00328 **
## GDP           0.037562   0.005851   6.419 2.82e-07 ***
## DJCA         -0.010574   0.007083  -1.493  0.14495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.43 on 33 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.9661
## F-statistic: 499.7 on 2 and 33 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE, data=DF_indicators)
summary(model2)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE, data = DF_indicators)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.865  -3.693  -1.033   3.978  17.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.52624   49.89618  -0.111    0.912
## GDP          0.04086    0.00204  20.026  < 2e-16 ***
## UNRATE      13.82381    2.24317   6.163 5.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.348 on 33 degrees of freedom
## Multiple R-squared:  0.9841, Adjusted R-squared:  0.9832
## F-statistic:  1024 on 2 and 33 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + HOUST, data=DF_indicators)
summary(model3)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + HOUST, data = DF_indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.734  -6.178  -0.912   5.764  35.614
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 254.963578  24.932722  10.226 9.22e-12 ***
## GDP           0.033336   0.002336  14.270 1.13e-15 ***
## HOUST        -0.013550   0.006645  -2.039   0.0495 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.16 on 33 degrees of freedom
## Multiple R-squared:  0.9697, Adjusted R-squared:  0.9679
## F-statistic:   528 on 2 and 33 DF,  p-value: < 2.2e-16
```

Each second predictor (DJCA, UNRATE, HOUST) has a p-value < 0.15, and is thus a candidate to enter the model. We choose to proceed with the one with the lowest p-value, which is UNRATE. Note that our p-value for GDP remains below 0.15, so we proceed with the MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE model.

3. Fit each three-predictor model that includes GDP and UNRATE.

```
model1 <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA, data=DF_indicators)
summary(model1)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA, data = DF_indicators)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -13.6477  -3.3376  -0.0371   2.9161  17.6972
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.818877  61.692482  -1.002    0.324
## GDP           0.046589   0.004310  10.810 3.26e-12 ***
## UNRATE       13.470870   2.214328   6.084 8.53e-07 ***
## DJCA         -0.007391   0.004926  -1.501    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.212 on 32 degrees of freedom
## Multiple R-squared:  0.9852, Adjusted R-squared:  0.9838
## F-statistic:   709 on 3 and 32 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + HOUST, data=DF_indicators)
summary(model2)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + HOUST, data = DF_indicators)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8922  -4.0787  -0.2048   4.0326  17.8333
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.155789  53.014829  -0.380    0.706
## GDP           0.040389   0.002124  19.012  < 2e-16 ***
## UNRATE       15.147181   2.743078   5.522 4.35e-06 ***
## HOUST         0.004972   0.005880   0.846    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.38 on 32 degrees of freedom
## Multiple R-squared:  0.9845, Adjusted R-squared:  0.983
## F-statistic: 676.7 on 3 and 32 DF,  p-value: < 2.2e-16
```

Since the p-value for HOUST is greater than 0.15, it cannot be added to the model. We thus proceed with MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA.

4. Fit each four-predictor model that includes GDP, UNRATE, and DJCA.

```
model <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA + HOUST, data=DF_indicators)
summary(model)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA + HOUST,
##     data = DF_indicators)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3470  -4.1344  -0.1523   3.5106  18.1108
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68.095184  63.190012  -1.078    0.290
## GDP           0.045830   0.004533  10.110 2.47e-11 ***
## UNRATE       14.439271   2.758068   5.235 1.09e-05 ***
## DJCA         -0.006848   0.005058  -1.354    0.186
## HOUST         0.003541   0.005900   0.600    0.553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.285 on 31 degrees of freedom
## Multiple R-squared:  0.9853, Adjusted R-squared:  0.9835
## F-statistic: 521.2 on 4 and 31 DF,  p-value: < 2.2e-16
```

Note that the p-value for HOUST is greater than our alpha_to_enter, so we do not include it in the model. Thus, our final regression model is as follows:

```
model_final <- lm(MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA, data=DF_indicators)
summary(model_final)
```

```
##
## Call:
## lm(formula = MEDIAN_WEEKLY_WAGES ~ GDP + UNRATE + DJCA, data = DF_indicators)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6477  -3.3376  -0.0371   2.9161  17.6972
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.818877  61.692482  -1.002    0.324
## GDP           0.046589   0.004310  10.810 3.26e-12 ***
## UNRATE       13.470870   2.214328   6.084 8.53e-07 ***
## DJCA         -0.007391   0.004926  -1.501    0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.212 on 32 degrees of freedom
## Multiple R-squared:  0.9852, Adjusted R-squared:  0.9838
## F-statistic:   709 on 3 and 32 DF,  p-value: < 2.2e-16
```