Evanston Food Insec Model Dev

Adi Tyagi

8/14/2020

Load libraries and obtain data

We begin modelling with the Evanston only data.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

```
####################MODEL WITH THE EVANSTON ONLY NOW#############
fa_pm_evanston = read_csv("../Data/Modelling Data/fa_pm_evanston-model.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Tract = col_character(),
##   state = col_character(),
##   total_population = col_number(),
##   food_insecurity_num_2020 = col_number()
## )
```

```
## See spec(...) for full column specifications.
```

```
#convert column types
fa_pm_evanston$total_population = as.numeric(fa_pm_evanston$total_population)
fa_pm_evanston$food_insecurity_num_2020 =
as.numeric(fa_pm_evanston$food_insecurity_num_2020)
```

```
fa_pm_evanston$food_insecurity_num_2018 =
as.numeric(fa_pm_evanston$food_insecurity_num_2018)
```

Model Building

We can do so by iterate through variables in list. They are in descending order of correlation with food_insecurity_num_2020. At each step, we add a variable from the list and compute the adjusted r-sq. In the end, we pick the variable subset with the highest adj r-sq.

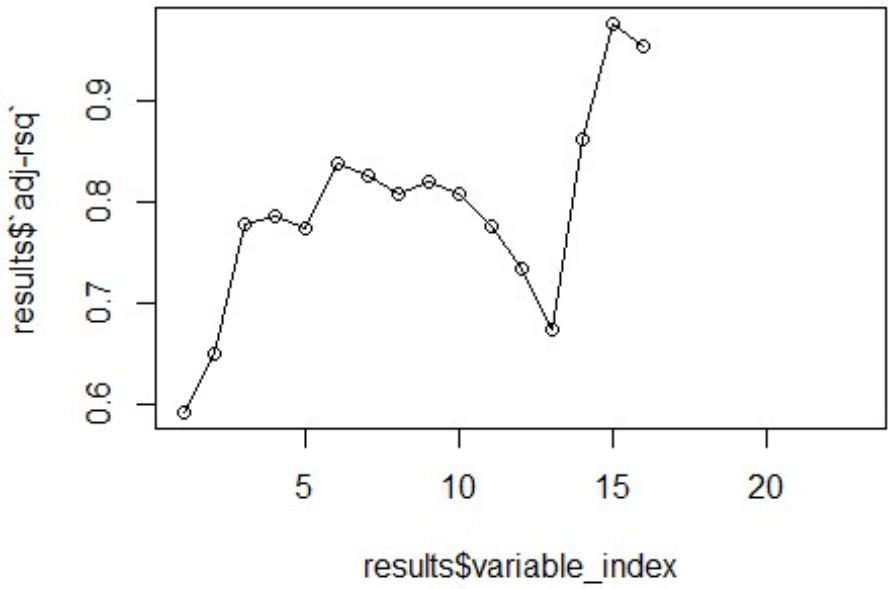
```
var_list = c( 'median_household_income',
'median_age', 'social_vulnerability_index',
'median_home_loan_amount', 'median_home_value', 'prop_nonwhite',
'prop_nonenglish_speaking', 'proportion_hispanic',
'thiel_racial_segregation_index', 'proportion_bachelors_degree',
'computer_access', 'prop_families_poverty', 'num_jobs',
'num_housing_units', 'life_expectancy', 'avg_travel_time_to_work',
'prop_students_in_public_school', 'median_leverage_ratio',
'proportion_disabled', 'avg_household_size', 'prop_men',
'local_census_tract', 'unemployment_change')

mod_set = c("food_insecurity_num_2020")
results = data.frame(index = c(), adj_rsq = c())
i = 0
for (var in var_list) {
  mod_set = append(mod_set, var)
  i = i + 1
  lin.mod = lm(food_insecurity_num_2020 ~ .,
               data = fa_pm_evanston %>% select(mod_set))
  adj_rsq = summary(lin.mod)$adj.r.squared
  results = rbind(results, list(i, adj_rsq))
}
names(results)[1] = "variable_index"
names(results)[2] = "adj-rsq"
results
```

```
##    variable_index   adj-rsq
## 1               1 0.5924345
## 2               2 0.6491627
## 3               3 0.7781963
## 4               4 0.7860126
## 5               5 0.7733248
## 6               6 0.8369056
## 7               7 0.8260374
## 8               8 0.8076374
## 9               9 0.8193072
## 10             10 0.8066524
## 11             11 0.7757305
## 12             12 0.7331380
## 13             13 0.6748509
```

```
## 14              14 0.8618924
## 15              15 0.9750775
## 16              16 0.9528979
## 17              17       NaN
## 18              18       NaN
## 19              19       NaN
## 20              20       NaN
## 21              21       NaN
## 22              22       NaN
## 23              23       NaN
```

Let's graph the results

```
#results %>% ggplot() + geom_point(aes(x = variable_index, y = adj_rsq))
plot(results$variable_index, results$`adj-rsq`, type = 'o')
```



Initial Subset

The variable subset with the highest adj-rsq has been identified as: - median_household_income - median_age - social_vulnerabillity_index - median_home_loan_amount - median_home_value - prop_nonwhite - prop_nonenglish_speaking - proportion_hispanic - thiel_racial_segregation_index - proportion_bachelors_degree - computer_access - prop_families_poverty - num_jobs - num_housing_units - life_expectancy

```
vars_list.subset = c( 'median_household_income',
                      'median_age', 'social_vulnerability_index',
                      'median_home_loan_amount', 'median_home_value',
'prop_nonwhite',
```

```
                      'prop_nonenglish_speaking', 'proportion_hispanic',
                      'thiel_racial_segregation_index',
'proportion_bachelors_degree',
                      'computer_access', 'prop_families_poverty', 'num_jobs',
                      'num_housing_units', 'life_expectancy',
'food_insecurity_num_2020')
mod.evanston.subset1 = lm(food_insecurity_num_2020 ~.,
                          data = fa_pm_evanston %>% select(vars_list.subset))
summary(mod.evanston.subset1)

##
## Call:
## lm(formula = food_insecurity_num_2020 ~ ., data = fa_pm_evanston %>%
##     select(vars_list.subset))
##
## Residuals:
##        1        2        3        4        5        6        7        8
##   0.8453  -9.7020  25.9527   8.3814 -22.4034  10.7062   4.2829   9.1808
##        9       10       11       12       13       14       15       16
##  -5.6700 -17.9284  -0.8304  -7.9435   0.3610  -6.0070  -2.0315 -11.4328
##       17       18
##   3.3852  20.8533
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.337e+03  9.023e+02  -1.482   0.2766
## median_household_income        4.513e-03  9.922e-04   4.548   0.0451 *
## median_age                    -3.233e+01  4.683e+00  -6.902   0.0204 *
## social_vulnerability_index     8.922e+02  2.047e+02   4.359   0.0488 *
## median_home_loan_amount        1.757e-03  4.461e-04   3.938   0.0589 .
## median_home_value             -1.053e-03  2.667e-04  -3.947   0.0586 .
## prop_nonwhite                  6.985e+00  3.308e+00   2.111   0.1691
## prop_nonenglish_speaking      -6.155e+00  5.413e+00  -1.137   0.3733
## proportion_hispanic            9.859e+00  4.635e+00   2.127   0.1673
## thiel_racial_segregation_index -1.530e+03  3.556e+02  -4.301   0.0500 .
## proportion_bachelors_degree    2.546e+01  7.922e+00   3.214   0.0847 .
## computer_access               -1.736e+01  5.081e+00  -3.416   0.0760 .
## prop_families_poverty         -3.309e+00  5.556e+00  -0.596   0.6119
## num_jobs                      -3.615e-02  1.170e-02  -3.090   0.0907 .
## num_housing_units              8.493e-02  2.199e-02   3.863   0.0610 .
## life_expectancy                3.863e+01  1.010e+01   3.824   0.0621 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36 on 2 degrees of freedom
## Multiple R-squared:  0.9971, Adjusted R-squared:  0.9751
## F-statistic: 45.34 on 15 and 2 DF,  p-value: 0.02178
```

Subset 2

We now retry the model with another subset, which is smaller and has only significant variables.

```
mod.evanston.subset2 = lm(food_insecurity_num_2020 ~
                            median_household_income +
                            median_age +
                            social_vulnerability_index +
                            median_home_loan_amount +
                            median_home_value +
                            thiel_racial_segregation_index +
                            proportion_bachelors_degree  +
                            computer_access +
                            num_jobs +
                            num_housing_units +
                            life_expectancy,
                          data = fa_pm_evanston)
summary(mod.evanston.subset2)

##
## Call:
## lm(formula = food_insecurity_num_2020 ~ median_household_income +
##     median_age + social_vulnerability_index + median_home_loan_amount +
##     median_home_value + thiel_racial_segregation_index +
proportion_bachelors_degree +
##     computer_access + num_jobs + num_housing_units + life_expectancy,
##     data = fa_pm_evanston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.395  -7.446   3.022  12.534  47.832
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.738e+02  7.859e+02   0.476  0.65114
## median_household_income         2.821e-03  1.009e-03   2.795  0.03137 *
## median_age                     -2.370e+01  4.965e+00  -4.773  0.00308 **
## social_vulnerability_index      7.552e+02  1.795e+02   4.207  0.00564 **
## median_home_loan_amount         1.219e-03  5.349e-04   2.280  0.06284 .
## median_home_value              -1.063e-03  3.428e-04  -3.102  0.02106 *
## thiel_racial_segregation_index -8.876e+02  2.707e+02  -3.279  0.01683 *
## proportion_bachelors_degree     4.526e+00  4.624e+00   0.979  0.36547
## computer_access                -1.121e+01  6.468e+00  -1.733  0.13385
## num_jobs                       -1.307e-02  8.839e-03  -1.479  0.18961
## num_housing_units               7.623e-02  2.710e-02   2.813  0.03062 *
## life_expectancy                 1.976e+01  1.014e+01   1.950  0.09909 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 50.81 on 6 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9504
## F-statistic: 30.58 on 11 and 6 DF,  p-value: 0.0002265
```

Subset 3

We can remove more variables that were not significant, and rebuild a smaller more parsimonious model.

```
mod.evanston.subset3 = lm(food_insecurity_num_2020 ~
                            median_household_income +
                            median_age +
                            social_vulnerability_index +
                            median_home_loan_amount +
                            median_home_value +
                            thiel_racial_segregation_index +
                            num_housing_units +
                            life_expectancy,
                          data = fa_pm_evanston)
summary(mod.evanston.subset3)
```

```
##
## Call:
## lm(formula = food_insecurity_num_2020 ~ median_household_income +
##     median_age + social_vulnerability_index + median_home_loan_amount +
##     median_home_value + thiel_racial_segregation_index + num_housing_units
+
##     life_expectancy, data = fa_pm_evanston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.413 -34.773   0.588  30.082  89.139
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.518e+02  5.813e+02   1.293 0.228094
## median_household_income        2.520e-03  1.068e-03   2.360 0.042635 *
## median_age                    -1.921e+01  3.886e+00  -4.943 0.000799 ***
## social_vulnerability_index     7.440e+02  1.267e+02   5.871 0.000237 ***
## median_home_loan_amount        6.841e-04  5.488e-04   1.246 0.244067
## median_home_value             -5.873e-04  3.124e-04  -1.880 0.092795 .
## thiel_racial_segregation_index -6.748e+02  2.849e+02  -2.368 0.042016 *
## num_housing_units              9.544e-02  2.741e-02   3.482 0.006912 **
## life_expectancy                4.428e-01  6.958e+00   0.064 0.950647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.92 on 9 degrees of freedom
## Multiple R-squared:  0.9647, Adjusted R-squared:  0.9332
## F-statistic: 30.71 on 8 and 9 DF,  p-value: 1.2e-05
```

Subset 4

We can remove more variables that were not significant, and rebuild a smaller more parsimonious model. This shall be our final model.

```
mod.evanston.subset4 = lm(food_insecurity_num_2020 ~
                            median_household_income +
                            median_age +
                            social_vulnerability_index +
                            median_home_value +
                            thiel_racial_segregation_index +
                            num_housing_units,
                          data = fa_pm_evanston)
summary(mod.evanston.subset4)
```

```
##
## Call:
## lm(formula = food_insecurity_num_2020 ~ median_household_income +
##     median_age + social_vulnerability_index + median_home_value +
##     thiel_racial_segregation_index + num_housing_units, data =
fa_pm_evanston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.642 -34.380  -4.397  20.455 120.089
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    8.648e+02  1.193e+02   7.248 1.65e-05 ***
## median_household_income         2.283e-03  1.008e-03   2.266 0.044649 *
## median_age                     -1.851e+01  3.727e+00  -4.967 0.000424 ***
## social_vulnerability_index      6.902e+02  1.154e+02   5.983 9.15e-05 ***
## median_home_value              -2.604e-04  1.370e-04  -1.901 0.083755 .
## thiel_racial_segregation_index -5.782e+02  2.629e+02  -2.200 0.050098 .
## num_housing_units               9.254e-02  2.413e-02   3.836 0.002767 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.69 on 11 degrees of freedom
## Multiple R-squared:  0.9572, Adjusted R-squared:  0.9338
## F-statistic: 40.96 on 6 and 11 DF,  p-value: 6.766e-07
```

Key Takeaways

In the end, we have created a 6 variable model with a 0.9338 adj. r-squared. A positive effect indicates that higher values of the variable are associated with higher values of food insecurity; A negative effect indicates that lower values of the variable are associated with higher varlues of food insecurity.

Overall, the takeaways are the following: - median_household_income: positive effect - median_age: negative effect - social_vulnerability_index: positive effect - median_home_value: negative effect - thiel_racial_segregation_index: negative effect - num_housing_units: positive effect