Baseline and Regression Models

12/07/2019

```
# load necessary packages

library(readr)
library(dplyr)
library(GGally)
library(ggplot2)
library(car)
```

```
us <- read_csv("https://raw.githubusercontent.com/kanam12/ieor142finalproject/master/us_suicides_merged_
#names(suicides)[9] <- "suicides_rate"

suicides <- us %>%  select(-age, - `country-year`, -country)

set.seed(377)


train.ids = sample(nrow(suicides), 0.70*nrow(suicides))
train = suicides[train.ids,]
test = suicides[-train.ids,]
```

Baseline Model

```
base_mod <- mean(suicides$`suicides/100k pop`)
```

Linear Regression

```
set.seed(377)

exp_mod <- lm(`suicides/100k pop` ~ ., data = train)

summary(exp_mod)
```

```
##
## Call:
## lm(formula = `suicides/100k pop` ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9676  -3.2243   0.1071   2.9260  21.1764
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -6.561e+02  1.330e+03  -0.493  0.62221
## year                       2.763e-01  7.131e-01   0.387  0.69873
## sexmale                    8.961e+00  9.126e+00   0.982  0.32713
## suicides_no                2.155e-03  3.863e-04   5.579 6.39e-08 ***
```

```
## population                -6.970e-07  8.892e-08  -7.839 1.39e-13 ***
## `HDI for year`             1.605e+02  1.736e+02   0.925  0.35600
## `gdp_for_year ($)`         2.110e-12  2.369e-12   0.891  0.37394
## `gdp_per_capita ($)`      -1.112e-03  8.393e-04  -1.325  0.18633
## generationBoomers          5.670e+00  1.378e+00   4.113 5.33e-05 ***
## generationSilent           3.280e+00  1.241e+00   2.643  0.00875 **
## generationG.I. Generation  1.108e+01  1.713e+00   6.469 5.33e-10 ***
## generationMillenials      -2.833e+00  1.360e+00  -2.083  0.03825 *
## generationGeneration Z    -6.841e+00  2.624e+00  -2.607  0.00969 **
## depression_percentage     -3.811e-01  3.736e+00  -0.102  0.91884
## drug_death_rate            1.284e-01  7.705e-02   1.666  0.09699 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.205 on 245 degrees of freedom
## Multiple R-squared:  0.8063, Adjusted R-squared:  0.7952
## F-statistic: 72.84 on 14 and 245 DF,  p-value: < 2.2e-16
```

```
vif(exp_mod) #- perfect multicolinearity
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## year                 283.510832  1       16.837780
## sex                  140.526937  1       11.854406
## suicides_no            7.524014  1        2.742994
## population             5.056416  1        2.248648
## `HDI for year`       108.411003  1       10.412060
## `gdp_for_year ($)`   688.789938  1       26.244808
## `gdp_per_capita ($)` 747.303383  1       27.336850
## generation             6.593310  5        1.207565
## depression_percentage 140.933912  1       11.871559
## drug_death_rate        7.158119  1        2.675466
```

```
#alias(exp_mod)
```

```
set.seed(377)
lin_mod <- lm(`suicides/100k pop` ~ .-`gdp_per_capita ($)`, data = train)

summary(lin_mod)
```

```
##
## Call:
## lm(formula = `suicides/100k pop` ~ . - `gdp_per_capita ($)`,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4011  -2.8929  -0.0201   3.1459  21.7702
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -8.033e+02  1.327e+03  -0.605  0.54555
## year                       4.072e-01  7.073e-01   0.576  0.56535
## sexmale                    9.896e+00  9.113e+00   1.086  0.27856
## suicides_no                2.207e-03  3.849e-04   5.735 2.85e-08 ***
## population                -7.068e-07  8.875e-08  -7.964 6.18e-14 ***
## `HDI for year`             1.158e+01  1.325e+02   0.087  0.93041
```

```
## `gdp_for_year ($)`        -6.552e-13  1.123e-12  -0.583  0.56026
## generationBoomers          5.617e+00  1.380e+00   4.070 6.33e-05 ***
## generationSilent           3.237e+00  1.242e+00   2.606  0.00973 **
## generationG.I. Generation  1.098e+01  1.714e+00   6.404 7.62e-10 ***
## generationMillenials      -2.814e+00  1.362e+00  -2.067  0.03983 *
## generationGeneration Z    -6.503e+00  2.615e+00  -2.486  0.01357 *
## depression_percentage      9.775e-02  3.724e+00   0.026  0.97908
## drug_death_rate            1.313e-01  7.713e-02   1.703  0.08988 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.215 on 246 degrees of freedom
## Multiple R-squared:  0.8049, Adjusted R-squared:  0.7946
## F-statistic: 78.06 on 13 and 246 DF,  p-value: < 2.2e-16
```

```
vif(lin_mod)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## year                  278.073514  1       16.675536
## sex                   139.686347  1       11.818898
## suicides_no             7.446130  1        2.728760
## population              5.021576  1        2.240887
## `HDI for year`         62.964063  1        7.934990
## `gdp_for_year ($)`    154.381650  1       12.425041
## generation              6.499695  5        1.205839
## depression_percentage 139.615852  1       11.815915
## drug_death_rate         7.152034  1        2.674329
```

```
set.seed(377)
lin_mod2 <- lm(`suicides/100k pop` ~ .-`gdp_per_capita ($)` - year, data = train)

summary(lin_mod2)
```

```
##
## Call:
## lm(formula = `suicides/100k pop` ~ . - `gdp_per_capita ($)` -
##     year, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1604  -3.1161   0.0812   2.8675  21.6604
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -4.094e+01  8.748e+01  -0.468   0.6402
## sexmale                    8.762e+00  8.885e+00   0.986   0.3250
## suicides_no                2.218e-03  3.840e-04   5.776 2.29e-08 ***
## population                -7.041e-07  8.850e-08  -7.955 6.44e-14 ***
## `HDI for year`             6.657e+01  9.169e+01   0.726   0.4685
## `gdp_for_year ($)`        -7.668e-14  5.014e-13  -0.153   0.8786
## generationBoomers          5.583e+00  1.377e+00   4.055 6.72e-05 ***
## generationSilent           3.218e+00  1.240e+00   2.595   0.0100 *
## generationG.I. Generation  1.102e+01  1.710e+00   6.443 6.07e-10 ***
## generationMillenials      -2.785e+00  1.359e+00  -2.049   0.0415 *
## generationGeneration Z    -6.641e+00  2.601e+00  -2.553   0.0113 *
```

```
## depression_percentage     -3.698e-01  3.630e+00  -0.102   0.9189
## drug_death_rate            1.282e-01  7.683e-02   1.668   0.0966 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.206 on 247 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.7951
## F-statistic: 84.77 on 12 and 247 DF,  p-value: < 2.2e-16
```

```
vif(lin_mod2)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## sex                   133.155360  1       11.539296
## suicides_no             7.429999  1        2.725803
## population              5.007900  1        2.237834
## `HDI for year`         30.238331  1        5.498939
## `gdp_for_year ($)`     30.843074  1        5.553654
## generation              6.356335  5        1.203153
## depression_percentage 132.977265  1       11.531577
## drug_death_rate         7.115567  1        2.667502
```

```
set.seed(377)
lin_mod3 <- lm(`suicides/100k pop` ~ .-`gdp_per_capita ($)` - year -sex, data = train)

summary(lin_mod3)
```

```
##
## Call:
## lm(formula = `suicides/100k pop` ~ . - `gdp_per_capita ($)` -
##     year - sex, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1977  -3.1137   0.0782   3.0937  21.5800
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.068e+01  7.009e+01   0.152  0.87905
## suicides_no                2.241e-03  3.832e-04   5.846 1.58e-08 ***
## population                -7.064e-07  8.847e-08  -7.985 5.25e-14 ***
## `HDI for year`             2.963e+01  8.368e+01   0.354  0.72355
## `gdp_for_year ($)`         1.219e-13  4.592e-13   0.265  0.79090
## generationBoomers          5.563e+00  1.377e+00   4.041 7.09e-05 ***
## generationSilent           3.331e+00  1.235e+00   2.697  0.00747 **
## generationG.I. Generation  1.105e+01  1.710e+00   6.462 5.43e-10 ***
## generationMillenials      -2.703e+00  1.356e+00  -1.993  0.04739 *
## generationGeneration Z    -6.535e+00  2.599e+00  -2.515  0.01254 *
## depression_percentage     -3.906e+00  5.635e-01  -6.931 3.60e-11 ***
## drug_death_rate            1.296e-01  7.681e-02   1.687  0.09293 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.206 on 248 degrees of freedom
## Multiple R-squared:  0.8039, Adjusted R-squared:  0.7952
## F-statistic:  92.4 on 11 and 248 DF,  p-value: < 2.2e-16
```

```r
vif(lin_mod3)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## suicides_no            7.402857  1        2.720819
## population             5.004313  1        2.237032
## `HDI for year`        25.192748  1        5.019238
## `gdp_for_year ($)`    25.868829  1        5.086141
## generation             6.284809  5        1.201792
## depression_percentage  3.205765  1        1.790465
## drug_death_rate        7.113163  1        2.667051
```

```r
# remove varibles that are not significant
set.seed(377)
lin_mod4 <- lm(`suicides/100k pop` ~ .-`gdp_for_year ($)` - `gdp_per_capita ($)` - year - sex, data = t

summary(lin_mod4)
```

```
##
## Call:
## lm(formula = `suicides/100k pop` ~ . - `gdp_for_year ($)` - `gdp_per_capita ($)` -
##     year - sex, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4038  -3.0834   0.1028   3.0491  21.6051
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -6.596e+00  2.597e+01  -0.254  0.79974
## suicides_no                2.251e-03  3.805e-04   5.915 1.09e-08 ***
## population                -7.099e-07  8.730e-08  -8.132 2.00e-14 ***
## `HDI for year`             5.045e+01  2.911e+01   1.733  0.08430 .
## generationBoomers          5.542e+00  1.372e+00   4.040 7.12e-05 ***
## generationSilent           3.358e+00  1.228e+00   2.733  0.00672 **
## generationG.I. Generation  1.106e+01  1.706e+00   6.484 4.76e-10 ***
## generationMillenials      -2.651e+00  1.340e+00  -1.979  0.04895 *
## generationGeneration Z    -6.356e+00  2.505e+00  -2.538  0.01177 *
## depression_percentage     -3.878e+00  5.527e-01  -7.016 2.15e-11 ***
## drug_death_rate            1.327e-01  7.576e-02   1.751  0.08110 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.194 on 249 degrees of freedom
## Multiple R-squared:  0.8038, Adjusted R-squared:  0.7959
## F-statistic:   102 on 10 and 249 DF,  p-value: < 2.2e-16
```

```r
lin_mod5 <- lm(`suicides/100k pop` ~ .-`gdp_for_year ($)`-`HDI for year`-`gdp_per_capita ($)`-year -sex
                                      data = train)

summary(lin_mod5)
```

```
##
## Call:
## lm(formula = `suicides/100k pop` ~ . - `gdp_for_year ($)` - `HDI for year` -
##     `gdp_per_capita ($)` - year - sex - `HDI for year`, data = train)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.289  -3.002  -0.117   3.230  20.683
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.815e+01  2.857e+00  13.354  < 2e-16 ***
## suicides_no                 2.110e-03  3.733e-04   5.653 4.29e-08 ***
## population                 -7.623e-07  8.225e-08  -9.268  < 2e-16 ***
## generationBoomers           5.602e+00  1.377e+00   4.069 6.33e-05 ***
## generationSilent            3.905e+00  1.192e+00   3.276   0.0012 **
## generationG.I. Generation   9.855e+00  1.564e+00   6.302 1.32e-09 ***
## generationMillenials       -1.682e+00  1.222e+00  -1.376   0.1701
## generationGeneration Z     -4.237e+00  2.195e+00  -1.931   0.0547 .
## depression_percentage      -3.834e+00  5.543e-01  -6.917 3.85e-11 ***
## drug_death_rate             2.253e-01  5.393e-02   4.178 4.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.219 on 250 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7943
## F-statistic: 112.1 on 9 and 250 DF,  p-value: < 2.2e-16
```

```r
# OSR-sqaured of newest seasonal model
base_predictions <- rep(base_mod, nrow(test))

#**confirm if this is correct
base_SSE = sum((train$`suicides/100k pop` - rep(base_mod, nrow(train)))^2)
base_SST = sum((train$`suicides/100k pop` - mean(train$`suicides/100k pop`))^2)
base_R2 = 1 - base_SSE/base_SST

base_SSE = sum((test$`suicides/100k pop` - base_predictions)^2)
base_SST = sum((test$`suicides/100k pop` - mean(train$`suicides/100k pop`))^2)
base_OSR2 = 1 - base_SSE/base_SST

# this builds a vector of predicted values on the test set
lin_predictions <- predict(lin_mod5, newdata = test)

lin_SSE = sum((test$`suicides/100k pop` - lin_predictions)^2)
lin_SST = sum((test$`suicides/100k pop` - mean(train$`suicides/100k pop`))^2)
lin_OSR2 = 1 - lin_SSE/lin_SST

#####---------- need to compare change in OSR2

exp_predictions <- predict(exp_mod, newdata = test)

exp_SSE = sum((test$`suicides/100k pop` - exp_predictions)^2)
exp_SST = sum((test$`suicides/100k pop` - mean(train$`suicides/100k pop`))^2)
exp_OSR2 = 1 - exp_SSE/exp_SST
# # OSR-sqaured of the initial exploratory model
# exp_predictions <- predict(mod_exp, newdata = wrangler_test)
#
```

```r
# exp_SSE = sum((wrangler_test$WranglerSales - exp_predictions)^2)
# exp_SST = sum((wrangler_test$WranglerSales - mean(wrangler_train$WranglerSales))^2)
# exp_OSR2 = 1 - exp_SSE/exp_SST

# compare change in R-squared and OSR-squared between the two models

#**confirm if R^2 for baseline is correct
R2 <- c("base_R2" = base_R2, "exp_OR2" = summary(exp_mod)$r.squared, "lin_R2" = summary(lin_mod5)$r.squa
R2
```

```
##     base_R2     exp_OR2      lin_R2
## -0.00054829  0.80628055  0.80143721
```

```r
OSR2 <- c("base_OSR2" = base_OSR2, "exp_OSR2" = exp_OSR2, "lin_OSR2" = lin_OSR2)
OSR2
```

```
##   base_OSR2    exp_OSR2    lin_OSR2
## 0.003987337 0.741355507 0.735686902
```