SS154 Assignment 3

```r
# install.packages("janitor")
# install.packages('gridExtra')
# install.packages('cobalt')

library(cobalt)
```

```
##  cobalt (Version 4.5.4, Build Date: 2024-02-26)
```

```r
library(rgenoud)
```

```
## ##  rgenoud (Version 5.9-0.10, Build Date: 2023-12-13)
## ##  See http://sekhon.berkeley.edu/rgenoud for additional documentation.
## ##  Please cite software as:
## ##   Walter Mebane, Jr. and Jasjeet S. Sekhon. 2011.
## ##   ``Genetic Optimization Using Derivatives: The rgenoud package for R.''
## ##   Journal of Statistical Software, 42(11): 1-26.
## ##
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggplot2)
library(stats)
library(MatchIt)
```

```
##
## Attaching package: 'MatchIt'

## The following object is masked from 'package:cobalt':
##
##     lalonde
```

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
load("/Users/somto/Downloads/replication_data.RData", verbose=TRUE)
```

```
## Loading objects:
##   table
```

```r
head(table)
```

```
## # A tibble: 6 x 22
##    year state   statenam statenum  fips     pop farsfats farsvmt mv330 pp0514
##   <dbl> <chr>   <chr>       <dbl> <dbl>   <dbl>    <dbl>   <dbl> <dbl>  <dbl>
## 1  1970 Alabama AL              1     1 3454557       NA      NA  1297  0.208
## 2  1971 Alabama AL              1     1 3497349       NA      NA  1320  0.204
## 3  1972 Alabama AL              1     1 3540003       NA      NA   678  0.198
## 4  1973 Alabama AL              1     1 3580759       NA      NA  1376  0.193
## 5  1974 Alabama AL              1     1 3627778       NA      NA  1118  0.188
## 6  1975 Alabama AL              1     1 3680495      902      NA  1087  0.184
## # i 12 more variables: pp1519 <dbl>, pp2024 <dbl>, pp2529 <dbl>, pp3034 <dbl>,
## #   pp3544 <dbl>, pp4554 <dbl>, pp5564 <dbl>, pp6500 <dbl>, auto <dbl>,
## #   bus <dbl>, truck <dbl>, mtrcycl <dbl>
```

```r
filtered_data <- table %>%
  filter(year %in% c(2013))
dim(filtered_data)
```

```
## [1] 51 22
```

```r
filtered_data
```

```
## # A tibble: 51 x 22
##     year state      statenam statenum  fips    pop farsfats farsvmt mv330 pp0514
##    <dbl> <chr>      <chr>       <dbl> <dbl>  <dbl>    <dbl>   <dbl> <dbl>  <dbl>
##  1  2013 Alabama    AL              1     1 4.83e6      853   65046    NA 0.129
##  2  2013 Alaska     AK              2     2 7.35e5       51    4848    NA 0.140
##  3  2013 Arizona    AZ              3     4 6.63e6      849   60586    NA 0.138
##  4  2013 Arkansas   AR              4     5 2.96e6      498   33493    NA 0.135
##  5  2013 California CA              5     6 3.83e7     3107  329534    NA 0.133
##  6  2013 Colorado   CO              6     8 5.27e6      482   46968    NA 0.133
##  7  2013 Connectic~ CT              7     9 3.60e6      286   30941    NA 0.125
##  8  2013 Delaware   DE              8    10 9.26e5       99    9308    NA 0.123
##  9  2013 Dist of C~ DC              9    11 6.46e5       20    3527    NA 0.0850
## 10  2013 Florida    FL             10    12 1.96e7     2403  192702    NA 0.115
## # i 41 more rows
## # i 12 more variables: pp1519 <dbl>, pp2024 <dbl>, pp2529 <dbl>, pp3034 <dbl>,
## #   pp3544 <dbl>, pp4554 <dbl>, pp5564 <dbl>, pp6500 <dbl>, auto <dbl>,
## #   bus <dbl>, truck <dbl>, mtrcycl <dbl>
```

```r
# get averages of youth accident percentages

av_dataset <- filtered_data %>%
  group_by(state) %>%
  mutate(average_percentage = mean(c(pp1519, pp2024), na.rm = TRUE))
head(av_dataset)
```

```
## # A tibble: 6 x 23
## # Groups:   state [6]
##    year state      statenam statenum  fips     pop farsfats farsvmt mv330 pp0514
##   <dbl> <chr>      <chr>       <dbl> <dbl>   <dbl>    <dbl>   <dbl> <dbl>  <dbl>
## 1  2013 Alabama    AL              1     1  4.83e6      853   65046    NA  0.129
## 2  2013 Alaska     AK              2     2  7.35e5       51    4848    NA  0.140
## 3  2013 Arizona    AZ              3     4  6.63e6      849   60586    NA  0.138
## 4  2013 Arkansas   AR              4     5  2.96e6      498   33493    NA  0.135
## 5  2013 California CA              5     6  3.83e7     3107  329534    NA  0.133
## 6  2013 Colorado   CO              6     8  5.27e6      482   46968    NA  0.133
## # i 13 more variables: pp1519 <dbl>, pp2024 <dbl>, pp2529 <dbl>, pp3034 <dbl>,
## #   pp3544 <dbl>, pp4554 <dbl>, pp5564 <dbl>, pp6500 <dbl>, auto <dbl>,
## #   bus <dbl>, truck <dbl>, mtrcycl <dbl>, average_percentage <dbl>
```

```r
needed <- av_dataset %>%
  select(state, statenam, pp1519, pp2024, average_percentage)
# dim(needed)
head(needed)
```

```
## # A tibble: 6 x 5
## # Groups:   state [6]
##   state      statenam pp1519 pp2024 average_percentage
##   <chr>      <chr>     <dbl>  <dbl>              <dbl>
## 1 Alabama    AL       0.0660 0.0740             0.0700
## 2 Alaska     AK       0.0670 0.0860             0.0765
## 3 Arizona    AZ       0.0680 0.0730             0.0705
## 4 Arkansas   AR       0.0660 0.0710             0.0685
## 5 California CA       0.0690 0.0760             0.0725
## 6 Colorado   CO       0.0640 0.0710             0.0675
```

```
write.csv(needed, file = "marijuana_x_accidents.csv", row.names = FALSE)
getwd()
```

```
## [1] "/Users/somto"
```

```
# using augmented data from Google sheets

data <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vQSblNQnsUppe-H4FrbliXT-nuUIy5RTJq-bO

names(data)
```

```
##  [1] "state"
##  [2] "statenam"
##  [3] "marijuana..medical.or.recreational..legalized.by.2011"
##  [4] "X2013_pp1519"
##  [5] "X2013_pp2024"
##  [6] "X2013_average_percentage"
##  [7] "X2013_public_transport_percentage"
##  [8] "X2013_unemployment_rate"
##  [9] "X2013_gdp"
## [10] "X2013_pop"
## [11] "X2013_per_capita_gdp"
```

```
data <- data %>% clean_names()
names(data)
```

```
##  [1] "state"
##  [2] "statenam"
##  [3] "marijuana_medical_or_recreational_legalized_by_2011"
##  [4] "x2013_pp1519"
##  [5] "x2013_pp2024"
##  [6] "x2013_average_percentage"
##  [7] "x2013_public_transport_percentage"
##  [8] "x2013_unemployment_rate"
##  [9] "x2013_gdp"
## [10] "x2013_pop"
## [11] "x2013_per_capita_gdp"
```

```
head(data)
```

```
##        state statenam marijuana_medical_or_recreational_legalized_by_2011
## 1    Alabama       AL                                                   0
## 2     Alaska       AK                                                   1
## 3    Arizona       AZ                                                   1
## 4   Arkansas       AR                                                   0
## 5 California       CA                                                   1
## 6   Colorado       CO                                                   1
##   x2013_pp1519 x2013_pp2024 x2013_average_percentage
## 1        0.066        0.074                   0.0700
## 2        0.067        0.086                   0.0765
## 3        0.068        0.073                   0.0705
## 4        0.066        0.071                   0.0685
## 5        0.069        0.076                   0.0725
## 6        0.064        0.071                   0.0675
##   x2013_public_transport_percentage x2013_unemployment_rate x2013_gdp x2013_pop
## 1                            0.0048                   0.072   180,727   4833722
## 2                            0.0184                   0.062    51,542    735132
## 3                            0.0236                   0.079   261,924   6626624
## 4                            0.0047                   0.072   115,745   2959373
## 5                            0.0531                   0.094 2,050,693  38332521
## 6                            0.0335                   0.071   273,721   5268367
##   x2013_per_capita_gdp
## 1             37388.79
## 2             70112.58
## 3             39526.01
## 4             39111.33
## 5             53497.47
## 6             51955.57
```

Linear Regression Analysis

```
# linear regression models (no matching)

model1 = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             x2013_per_capita_gdp,
           data=data)

model2 = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             log(x2013_per_capita_gdp),
           data=data)

model3 = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             log(x2013_per_capita_gdp) +
             x2013_unemployment_rate:log(x2013_per_capita_gdp),
           data=data)

model4 = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             x2013_per_capita_gdp +
             x2013_unemployment_rate:x2013_per_capita_gdp,
           data=data)
```

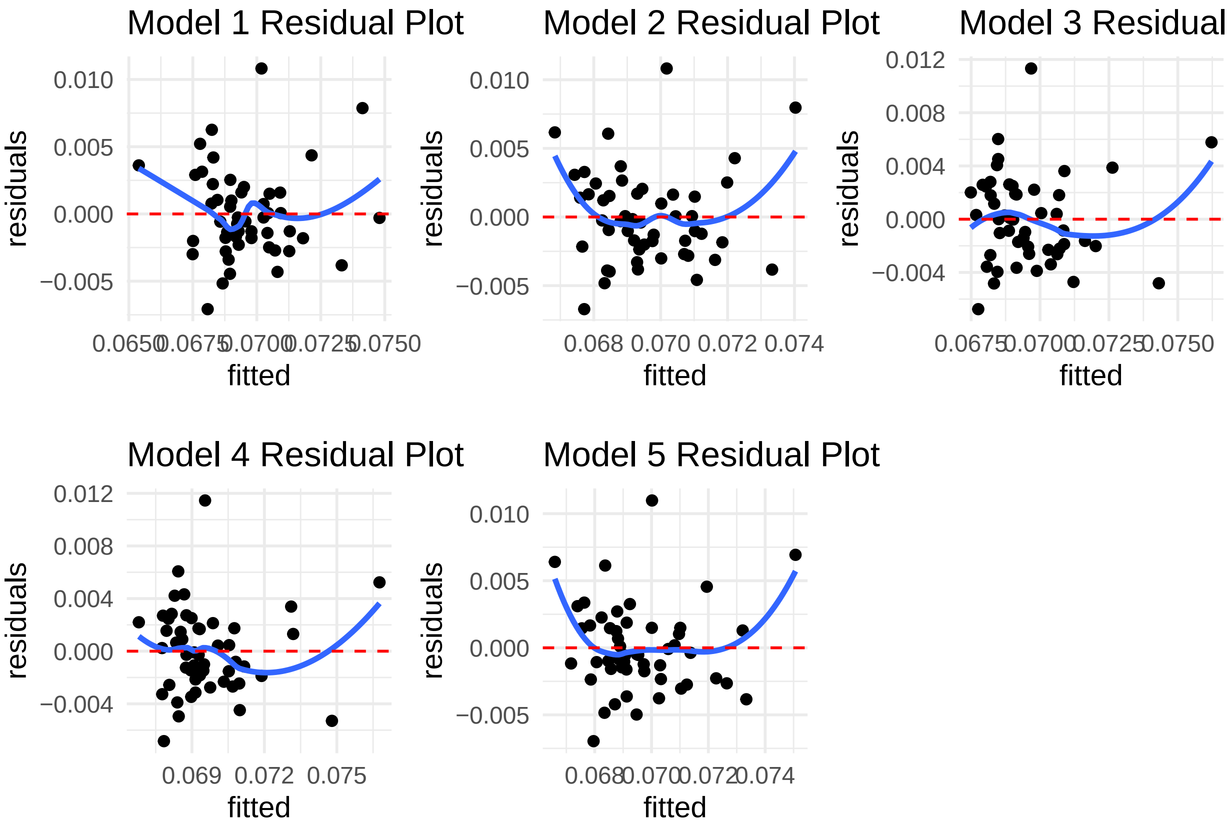
```r
model5 = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             log(x2013_per_capita_gdp) +
             x2013_public_transport_percentage:x2013_unemployment_rate,
           data=data)
```

```r
# stargazer(model1, model2, model3, model4, model5, type='html', out='ols_models_comparison.html')
stargazer(model1, model2, model3, model4, model5, type='text')
```

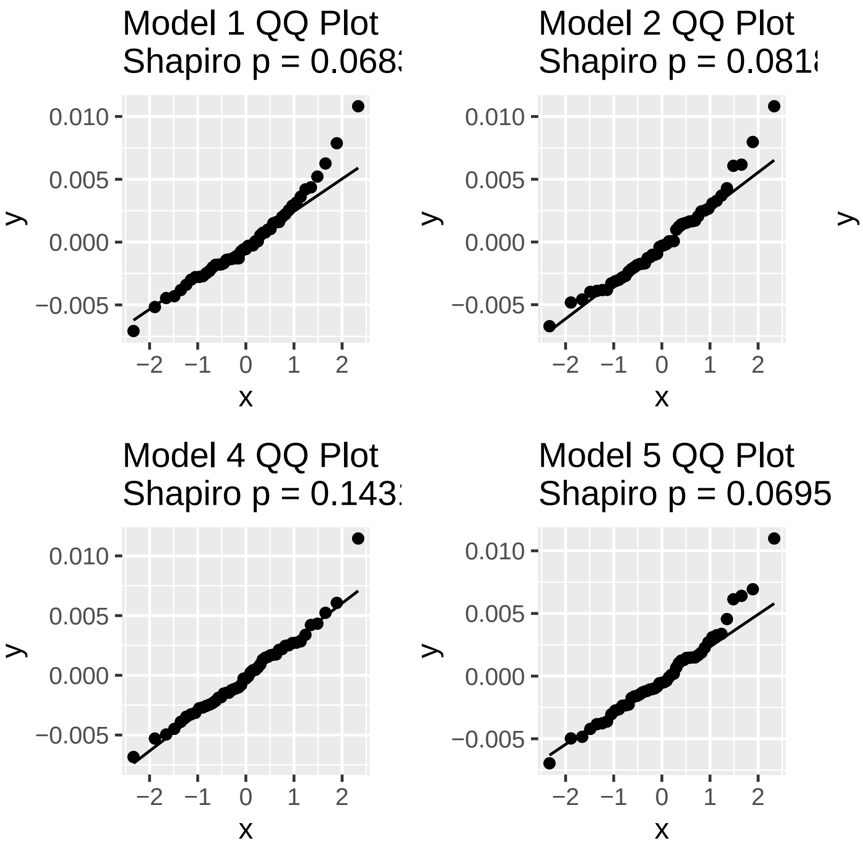
```
##
## ====================================================================================================
##                                                                                                     D
##                                                            ------------------------------------------
##                                                                                                   x20
##                                                                    (1)                  (2)
## ----------------------------------------------------------------------------------------------------
## marijuana_medical_or_recreational_legalized_by_2011              -0.001              -0.001
##                                                                 (0.001)              (0.001)
##
## x2013_public_transport_percentage                               -0.024*              -0.016
##                                                                 (0.014)              (0.013)
##
## x2013_unemployment_rate                                          -0.027              -0.026
##                                                                 (0.035)              (0.036)
##
## x2013_per_capita_gdp                                           0.00000***
##                                                                (0.00000)
##
## log(x2013_per_capita_gdp)                                                            0.007**
##                                                                                      (0.003)
##
## x2013_unemployment_rate:log(x2013_per_capita_gdp)
##
##
## x2013_unemployment_rate:x2013_per_capita_gdp
##
##
## x2013_public_transport_percentage:x2013_unemployment_rate
##
##
## Constant                                                        0.066***             -0.008
##                                                                 (0.004)              (0.036)
##
## ----------------------------------------------------------------------------------------------------
## Observations                                                       51                   51
## R2                                                               0.206                0.172
## Adjusted R2                                                      0.137                0.100
## Residual Std. Error                                         0.003 (df = 46)      0.004 (df = 46)
## F Statistic                                               2.987** (df = 4; 46) 2.393* (df = 4; 46) 2
## ====================================================================================================
## Note:
```

```
# residual plot to check linearity and homoskedasticity assumptions

# Function to calculate Goldfeld-Quandt test p-value
gq_test_p <- function(model) {
  gq_test_result <- gqtest(model)
  p_value <- gq_test_result$p.value
  return(p_value)
}

# Create residual plots for multiple models
residual_plots <- lapply(1:5, function(i) {
  model <- get(paste0("model", i))

  ggplot(data.frame(residuals = residuals(model), fitted = fitted(model)), aes(x = fitted, y = residuals
    geom_point() + geom_smooth(method = "loess", se=FALSE) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
    ggtitle(paste("Model", i, "Residual Plot")) +
    theme_minimal() +
    theme(aspect.ratio = 1)
})

# Arrange plots in a grid
grid.arrange(grobs = residual_plots, ncol = 3)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



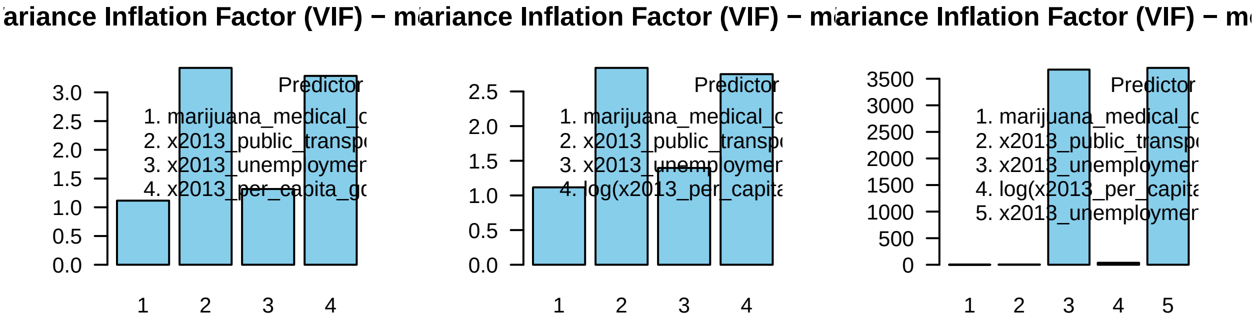
```
# QQ Plot - to check normality of errors assumption

# Function to calculate p-value from Shapiro-Wilk test
get_shapiro_p <- function(model) {
  shapiro_result <- shapiro.test(residuals(model))
  p_value <- shapiro_result$p.value
  return(p_value)
}

# Create QQ plots for multiple models
qqplots <- lapply(1:5, function(i) {
  model <- get(paste0("model", i))
  ggplot(data.frame(residuals = residuals(model)), aes(sample = residuals)) +
    stat_qq() +
    stat_qq_line() +
    ggtitle(paste("Model", i, "QQ Plot \nShapiro p =", round(get_shapiro_p(model), 4))) +
    theme(aspect.ratio = 1)  # Set aspect ratio to make plots more square
})

# Arrange plots in a grid
grid.arrange(grobs = qqplots, ncol = 3, main = "QQ Plots for Models 1 through 5")
```

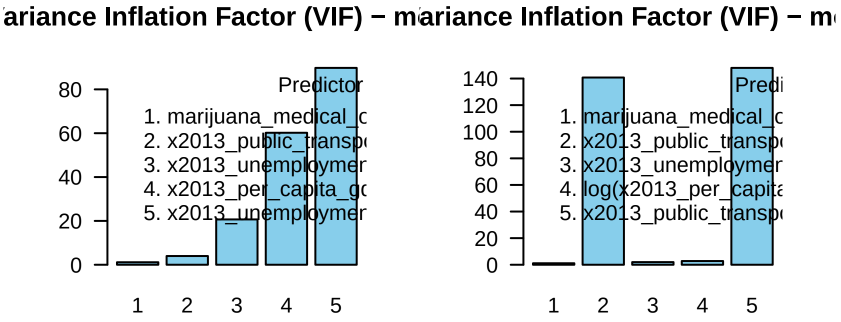


```r
# VIF plot - to check no multicollinearity assumption

# Set up multi-panel layout
par(mfrow = c(2, 3))  # 2 rows and 3 columns grid

# List of model names
model_names <- paste0("model", 1:5)

# Loop over each model
for (model_name in model_names) {
  # Get predictor names
  predictor_names <- names(coef(get(model_name)))[-1]

  # Create a numbered vector for the x-axis
  x_labels <- 1:length(predictor_names)

  # Plot with numbered labels
  barplot(vif(get(model_name)), col = "skyblue", main = paste("Variance Inflation Factor (VIF) -", model
          names.arg = x_labels, las = 1)

  # Create modified legend labels
  legend_labels <- paste0(x_labels, ". ", predictor_names) # Combine number and name

  # Create and position the legend
  legend("topleft", legend = legend_labels,
         title = "Predictor Number & Name",
         col = "skyblue", bty = "n", cex = 0.95)
}
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
# Reset the plotting layout
par(mfrow = c(1, 1))
```





Matching Analysis

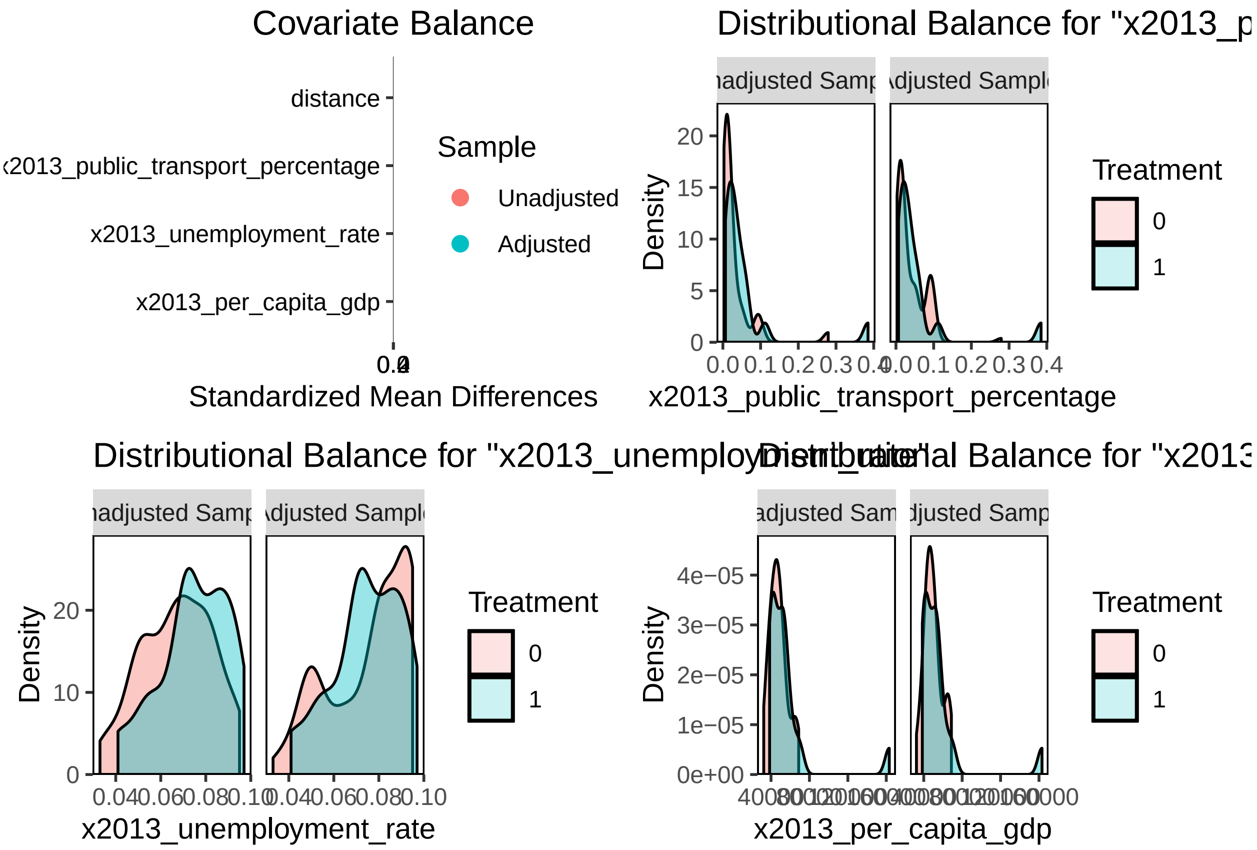
```
# full matching with propensity scores
m.out1 <- matchit(marijuana_medical_or_recreational_legalized_by_2011 ~
                    x2013_public_transport_percentage +
                    x2013_unemployment_rate +
                    x2013_per_capita_gdp,
                  data = data,
                  method = "full", distance = "glm")

# Checking balance after matching
summary(m.out1)
```

```
##
## Call:
## matchit(formula = marijuana_medical_or_recreational_legalized_by_2011 ~
##     x2013_public_transport_percentage + x2013_unemployment_rate +
##         x2013_per_capita_gdp, data = data, method = "full", distance = "glm")
```

```
##
## Summary of Balance for All Data:
##                                    Means Treated Means Control Std. Mean Diff.
## distance                                   0.4096        0.2952          0.5553
## x2013_public_transport_percentage          0.0552        0.0304          0.2795
## x2013_unemployment_rate                    0.0747        0.0671          0.4825
## x2013_per_capita_gdp                   56004.4134    47654.0888          0.2880
##                                    Var. Ratio eCDF Mean eCDF Max
## distance                               3.7144    0.1915   0.4412
## x2013_public_transport_percentage      3.0344    0.1887   0.4118
## x2013_unemployment_rate                0.9408    0.1292   0.2647
## x2013_per_capita_gdp                   9.5880    0.0952   0.2353
##
## Summary of Balance for Matched Data:
##                                    Means Treated Means Control Std. Mean Diff.
## distance                                   0.4096        0.3880          0.1047
## x2013_public_transport_percentage          0.0552        0.0382          0.1911
## x2013_unemployment_rate                    0.0747        0.0761         -0.0876
## x2013_per_capita_gdp                   56004.4134    49557.4793          0.2223
##                                    Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
## distance                               1.8664    0.0318   0.1176          0.1118
## x2013_public_transport_percentage      4.2015    0.0944   0.2608          0.4404
## x2013_unemployment_rate                0.7207    0.0766   0.2059          0.6349
## x2013_per_capita_gdp                   9.4694    0.0803   0.2137          0.3841
##
## Sample Sizes:
##               Control Treated
## All             34.        17
## Matched (ESS)   13.12      17
## Matched         34.        17
## Unmatched        0.         0
## Discarded        0.         0
```

```
m.out1_lp <- love.plot(m.out1)
m.out1_bp1 <- bal.plot(m.out1, "x2013_public_transport_percentage", which = "both")
m.out1_bp2 <- bal.plot(m.out1, "x2013_unemployment_rate", which = "both")
m.out1_bp3 <- bal.plot(m.out1, "x2013_per_capita_gdp", which = "both")
grid.arrange(m.out1_lp, m.out1_bp1, m.out1_bp2, m.out1_bp3, ncol = 2)
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
## No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
## No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```



```
m.genetic = matchit(marijuana_medical_or_recreational_legalized_by_2011 ~
                    x2013_public_transport_percentage +
                    x2013_unemployment_rate +
                    x2013_per_capita_gdp,
                  data = data, method="genetic", estimand = "ATT")
```

```
## Warning: (from Matching) The key tuning parameters for optimization were are
## all left at their default values.  The 'pop.size' option in particular should
## probably be increased for optimal results.  For details please see the help
## page and https://www.jsekhon.com
```

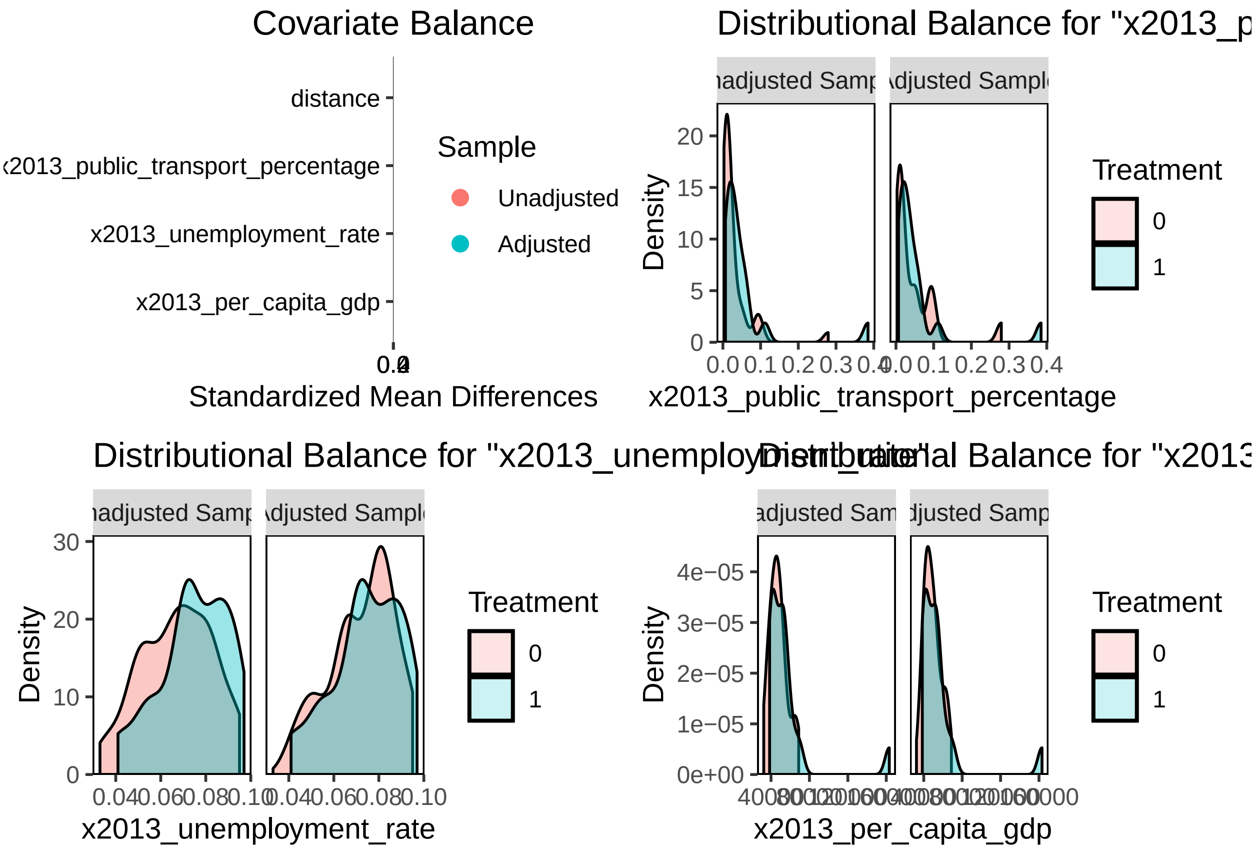
```
summary(m.genetic)
```

```
##
## Call:
## matchit(formula = marijuana_medical_or_recreational_legalized_by_2011 ~
##     x2013_public_transport_percentage + x2013_unemployment_rate +
##         x2013_per_capita_gdp, data = data, method = "genetic",
##     estimand = "ATT")
##
## Summary of Balance for All Data:
##                                   Means Treated Means Control Std. Mean Diff.
## distance                                 0.4096        0.2952          0.5553
## x2013_public_transport_percentage        0.0552        0.0304          0.2795
## x2013_unemployment_rate                  0.0747        0.0671          0.4825
## x2013_per_capita_gdp                 56004.4134    47654.0888          0.2880
##                                   Var. Ratio eCDF Mean eCDF Max
## distance                              3.7144    0.1915   0.4412
## x2013_public_transport_percentage     3.0344    0.1887   0.4118
## x2013_unemployment_rate               0.9408    0.1292   0.2647
## x2013_per_capita_gdp                  9.5880    0.0952   0.2353
##
## Summary of Balance for Matched Data:
##                                   Means Treated Means Control Std. Mean Diff.
## distance                                 0.4096        0.3354          0.3600
## x2013_public_transport_percentage        0.0552        0.0482          0.0796
## x2013_unemployment_rate                  0.0747        0.0728          0.1239
## x2013_per_capita_gdp                 56004.4134    48935.0393          0.2438
##                                   Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
## distance                              2.7762    0.1061   0.3529          0.4543
## x2013_public_transport_percentage     1.7079    0.1091   0.2353          0.2349
## x2013_unemployment_rate               1.1257    0.0553   0.1765          0.4167
## x2013_per_capita_gdp                 11.7594    0.0681   0.1765          0.3188
##
## Sample Sizes:
##           Control Treated
## All            34      17
## Matched        17      17
## Unmatched      17       0
## Discarded       0       0
```

```r
m.genetic_lp <- love.plot(m.genetic)
m.genetic_bp1 <- bal.plot(m.genetic, "x2013_public_transport_percentage", which = "both")
m.genetic_bp2 <- bal.plot(m.genetic, "x2013_unemployment_rate", which = "both")
m.genetic_bp3 <- bal.plot(m.genetic, "x2013_per_capita_gdp", which = "both")
grid.arrange(m.genetic_lp, m.genetic_bp1, m.genetic_bp2, m.genetic_bp3, ncol = 2)
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
## No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```



```
m1.data <- match.data(m.out1)
m.genetic.data <- match.data(m.genetic)
```

```
model2_full = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             log(x2013_per_capita_gdp),
           data=m1.data)

model2_genetic = lm(x2013_average_percentage ~
             marijuana_medical_or_recreational_legalized_by_2011 +
             x2013_public_transport_percentage +
             x2013_unemployment_rate +
             log(x2013_per_capita_gdp),
           data=m.genetic.data)

# stargazer(model2_full, model2_genetic, type='html', out='ols_model2_matching_comparison.html')
stargazer(model2_full, model2_genetic, type='text')
```

```
##
## ===========================================================================================
##                                                                 Dependent variable:
##                                                         --------------------------------------
##                                                                x2013_average_percentage
##                                                                 (1)                (2)
## -------------------------------------------------------------------------------------------
## marijuana_medical_or_recreational_legalized_by_2011       -0.001              -0.001
##                                                           (0.001)            (0.001)
##
## x2013_public_transport_percentage                         -0.016              -0.016
##                                                           (0.013)            (0.012)
##
## x2013_unemployment_rate                                   -0.026              -0.002
##                                                           (0.036)            (0.036)
##
## log(x2013_per_capita_gdp)                                 0.007**            0.008**
##                                                           (0.003)            (0.003)
##
## Constant                                                  -0.008              -0.013
##                                                           (0.036)            (0.037)
##
## ------------------------------------------------------------------------------------------
## Observations                                                51                  34
## R2                                                         0.172              0.168
## Adjusted R2                                                0.100              0.054
## Residual Std. Error                                   0.004 (df = 46)    0.003 (df = 29)
## F Statistic                                         2.393* (df = 4; 46) 1.469 (df = 4; 29)
## ==========================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```