3.3.1 - R: Influential Observations and Outliers

Stat 5100: Dr. Bean

Example: Data collected on 50 countries relevant to a cross-sectional study of a life-cycle savings hypothesis, which states that the response variable

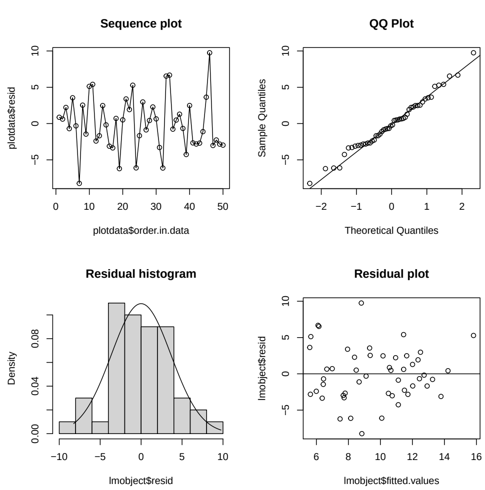
- SavRatio: aggregate personal saving divided by disposable income

can be explained by the following four predictor variables:

- AvIncome: per-capita disposable income, in USD (yearly average over decade)
- GrowRate: percentage growth rate in per-capita disposable income (over decade)
- PopU15: percentage of the population less than 15 years old (yearly average over decade)
- PopO75: percentage of the population over 75 years old (yearly average over decade)

The decade is 1960-1970. These data are published in section 2.2 of Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (1980) by Belsley, Kuh, and Welsch (limited excerpt available through Google books).

```
# Load in and take a look at the data
library(stat5100)
data(savings)

# Create a regression model to predict SavRatio
savings_lm <- lm(SavingsRatio ~ PctPopU15 + PctPop075 + AverageIncome + GrowthRate,
                 data = savings)

# Look at some basic visual assumptions
stat5100::visual_assumptions(savings_lm)
```



```r
# Numerical assumptions
stat5100::brown_forsythe_lm(savings_lm)
```

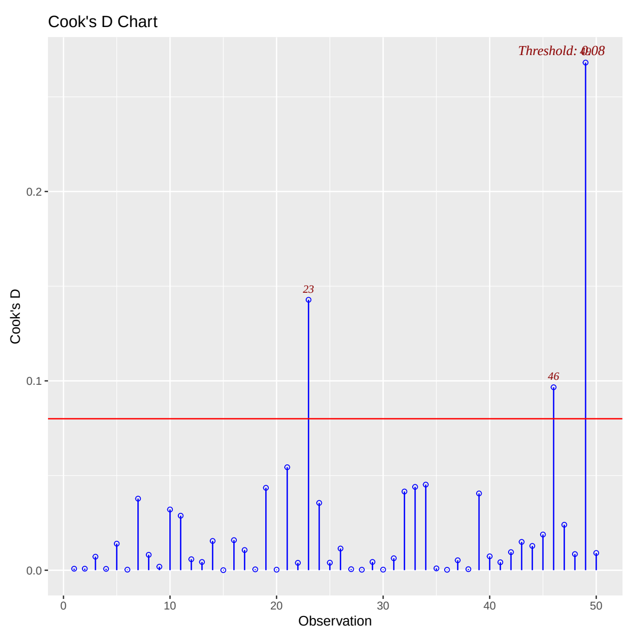
```
## [1] "Brown-forsythe test for constant variance in the residuals:"
## [1] "T-statistic: 1.9704, p-value: 0.0546"
```

```r
stat5100::cor_normality_lm(savings_lm)
```

```
## Correlation test of normality:
##                    resid expected_norm
## resid         1.0000000     0.9925168
## expected_norm 0.9925168     1.0000000
##
## Total observations: 50
## Make sure to consult with table B.6 for your final result.
```

Look at some diagnostics for influential observations and outliers

```r
# Cook's D Chart
olsrr::ols_plot_cooksd_chart(savings_lm)
```

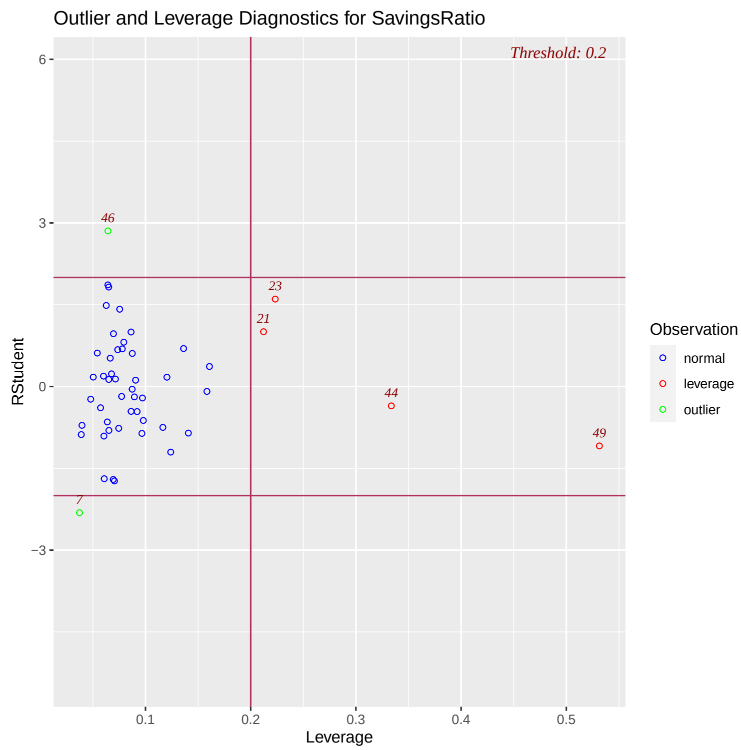



```r
# The output above for Cook's D doesn't tell us which names belong to the numbers
# in the graph. We can find them by indexing the country vector inside savings:
savings$Country[c(23, 46, 49)]
```

```
## [1] "Japan"  "Zambia" "Libya"
```

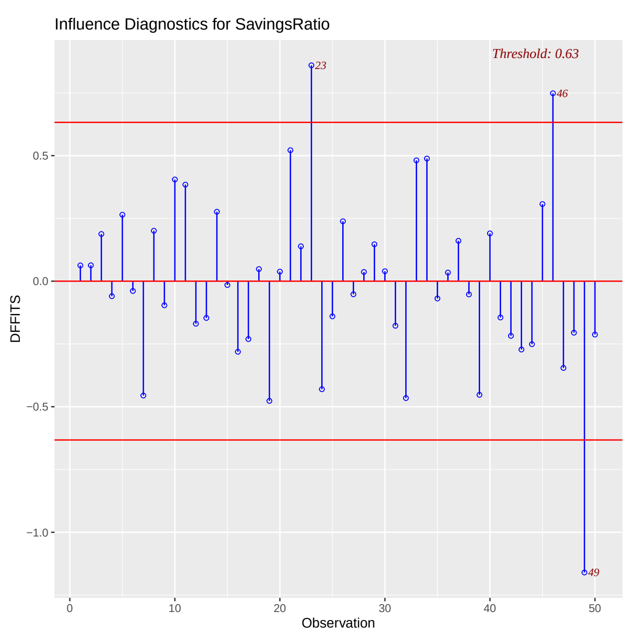
```r
# Which tells us that countries Japan, Zambia, and Libya strongly influenced
# the fitted values of the model.
```

```r
# Outlier and Leverage Diagnostics
olsrr::ols_plot_resid_lev(savings_lm)
```

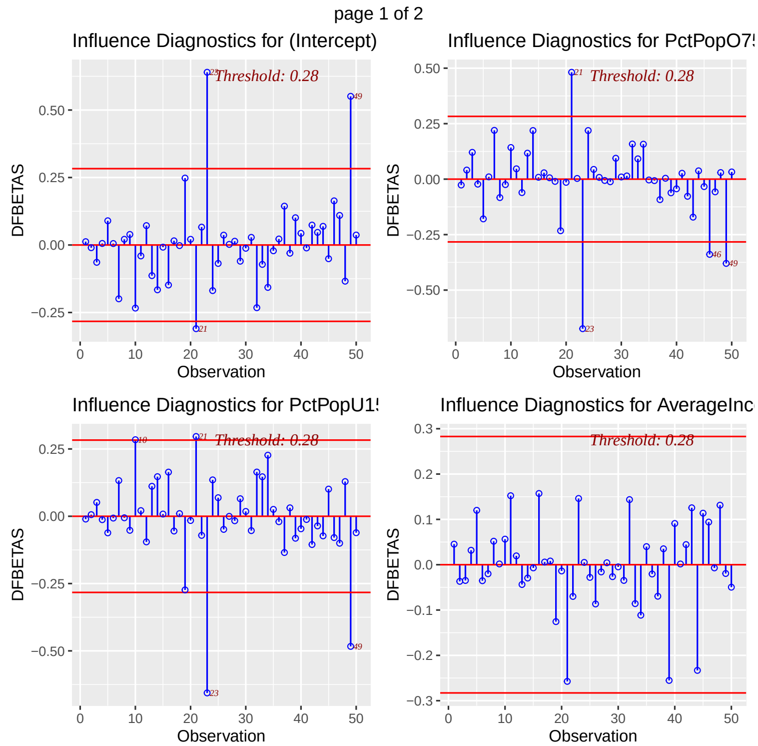


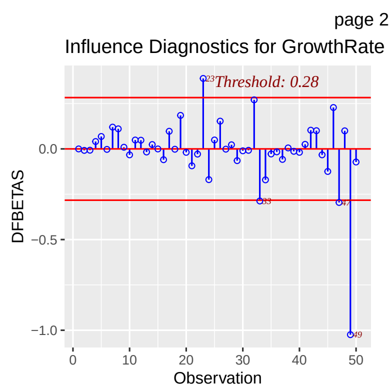

```r
# Once again we can find the names by indexing:
savings$Country[c(7, 46)] # Outliers

## [1] "Chile"  "Zambia"

savings$Country[c(23, 21, 44, 49)] # Leverage

## [1] "Japan"         "Ireland"       "United States" "Libya"
```

```r
# DFFITs plot:
olsrr::ols_plot_dffits(savings_lm)
```



```r
# DFBETAs panel:
olsrr::ols_plot_dfbetas(savings_lm)
```



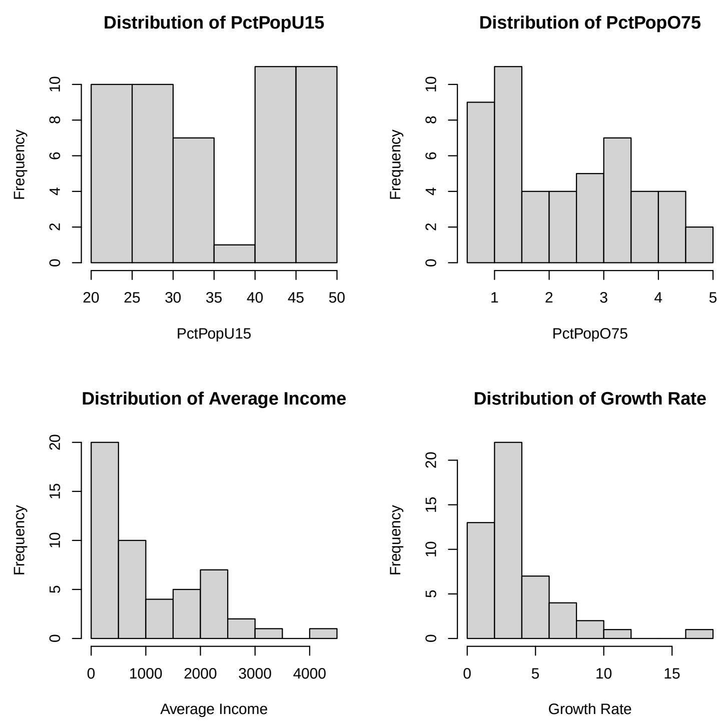


Alternative thresholds for influential observations and outlier diagnostics

```r
p <- 5 # Number of beta parameters, including intercept
n <- 50 # Sample size

cooks_d_simple <- 4 / n
cooks_d_10 <- qf(0.10, p, n-p)
cooks_d_20 <- qf(0.20, p, n-p)
cooks_d_50 <- qf(0.50, p, n-p)

rstudent_95 <- qt(1 - 0.05/2, n-p)
rstudent_95_bonf <- qt(1 - 0.05/(2*n), n-p)

leverage_2 <- 2 * p/n
leverage_3 <- 3 * p/n

# If we had n less than 30, then we should set both DFBETAS and DFFITS to 1.
DFBETAS = 2/(n^0.5)
DFFITS = 2*(p/n)^0.5

# Look at all the alternative thresholds:
thresholds <- data.frame(cooks_d_simple, cooks_d_10, cooks_d_20, cooks_d_50, rstudent_95,
                rstudent_95_bonf, leverage_2, leverage_3, DFBETAS, DFFITS)
thresholds

##   cooks_d_simple cooks_d_10 cooks_d_20 cooks_d_50 rstudent_95 rstudent_95_bonf
## 1           0.08  0.3172927   0.465266  0.8834915    2.014103         3.520251
##   leverage_2 leverage_3   DFBETAS    DFFITS
## 1        0.2        0.3 0.2828427 0.6324555
```

Look closely at the distribution of predictors and the suspect observations

```
# Tile 4 histograms together of the four predictors
par(mfrow = c(2, 2))
hist(savings$PctPopU15, main = "Distribution of PctPopU15", xlab = "PctPopU15")
hist(savings$PctPop075, main = "Distribution of PctPop075", xlab = "PctPop075")
hist(savings$AverageIncome, main = "Distribution of Average Income",
     xlab = "Average Income")
hist(savings$GrowthRate, main = "Distribution of Growth Rate",
     xlab = "Growth Rate")
```

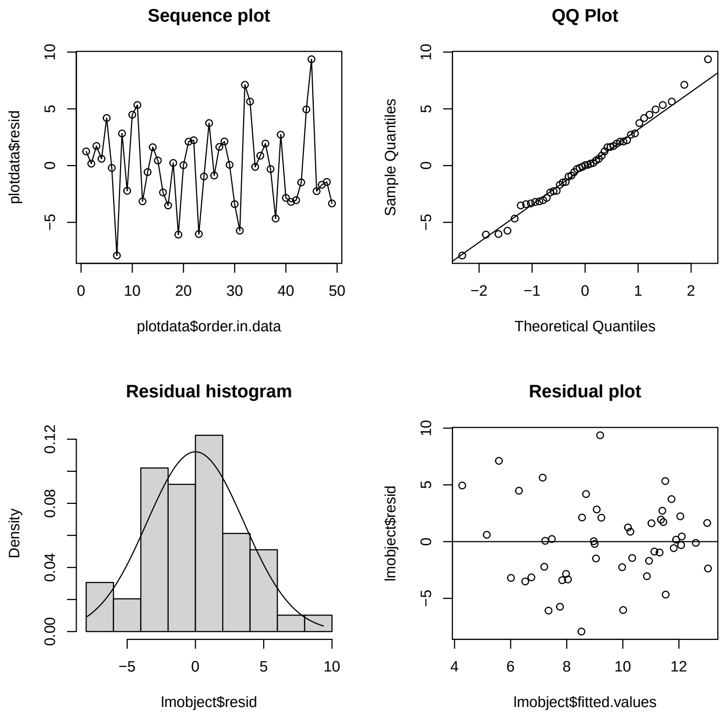



```
# Reset plot to just one graph per plot
par(mfrow = c(1, 1))

# Look at the suspect observations (previously identified influential points
# and outliers)

suspect_observations <- savings[savings$Country %in% c("Ireland", "Japan",
                                                        "United States", "Libya",
                                                        "Zambia"), ]
suspect_observations

##          Country SavingsRatio PctPopU15 PctPop075 AverageIncome GrowthRate
## 21       Ireland        11.34     31.16      4.19       1139.95       2.99
## 23         Japan        21.10     27.01      1.91       1257.28       8.21
## 44 United States         7.56     29.81      3.43       4001.89       2.45
## 46        Zambia        18.56     45.25      0.56        138.33       5.14
## 49         Libya         8.89     43.69      2.07        123.58      16.71
```

What are some possible remedial measures for this data?

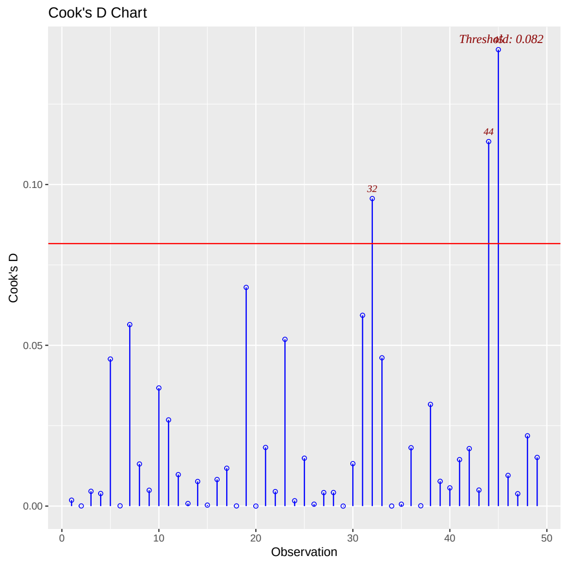
1. Drop Japan
   - PopU15 and PopO75 don't match the profile (influential but not outliers)
2. Take a log transform of AverageIncome and GrowthRate
   - Both distributions are skewed right
   - The extreme observations in each is a suspect observation (United States for AverageIncome, and Libya for GrowthRate)

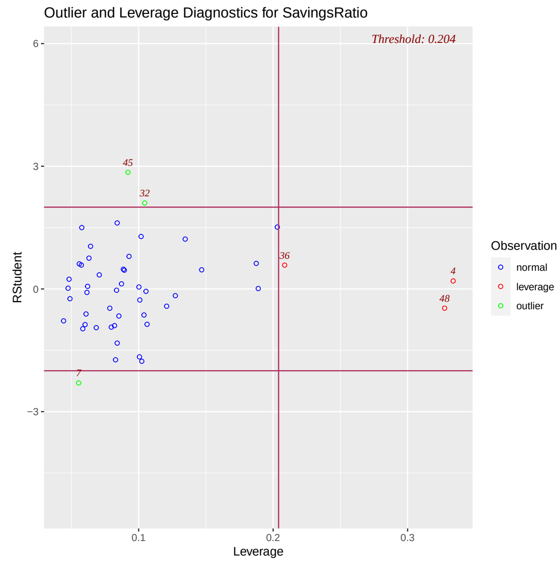
```
# Create new dataset, fit regression model, and then check assumptions
new_savings <- savings
new_savings <- new_savings[new_savings$Country != "Japan", ]
new_savings <- cbind(new_savings, logAverageIncome = log(new_savings$AverageIncome),
                     logGrowthRate = log(new_savings$GrowthRate))

new_savings_lm <- lm(SavingsRatio ~ PctPopU15 + PctPop075 + logAverageIncome +
                       logGrowthRate, data = new_savings)

summary(new_savings_lm)

##
## Call:
## lm(formula = SavingsRatio ~ PctPopU15 + PctPop075 + logAverageIncome +
##     logGrowthRate, data = new_savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9240 -2.3669  0.0355  2.1058  9.3683
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       26.2512    10.5263   2.494   0.0165 *
## PctPopU15         -0.3384     0.1579  -2.143   0.0377 *
## PctPop075         -0.6856     1.1357  -0.604   0.5492
## logAverageIncome  -0.7186     0.9749  -0.737   0.4650
## logGrowthRate      1.3304     0.7253   1.834   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.716 on 44 degrees of freedom
## Multiple R-squared:  0.2855,Adjusted R-squared:  0.2206
## F-statistic: 4.396 on 4 and 44 DF,  p-value: 0.004465

stat5100::visual_assumptions(new_savings_lm)
```



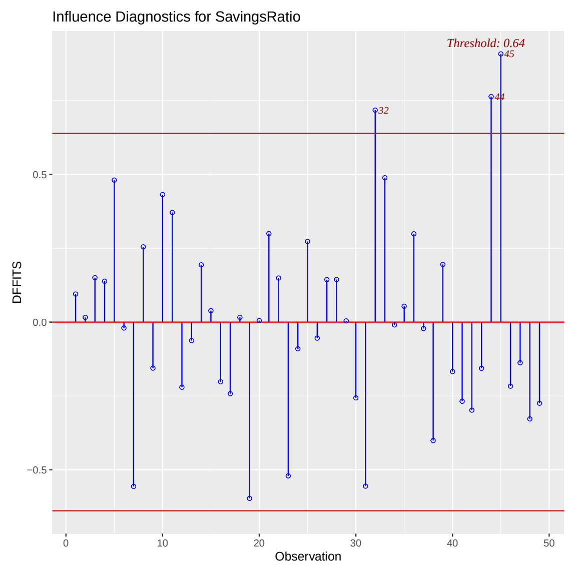

```
# Numerical assumptions
stat5100::brown_forsythe_lm(new_savings_lm)

## [1] "Brown-forsythe test for constant variance in the residuals:"
## [1] "T-statistic: 2.4334, p-value: 0.0188"

stat5100::cor_normality_lm(new_savings_lm)

## Correlation test of normality:
##                   resid expected_norm
## resid         1.0000000     0.9951555
## expected_norm 0.9951555     1.0000000
##
## Total observations: 49
## Make sure to consult with table B.6 for your final result.
```

```
# Check a few influential observation diagnostics
# -------------------------------
# Cook's D and Residual / Leverage plot
olsrr::ols_plot_cooksd_chart(new_savings_lm)
olsrr::ols_plot_resid_lev(new_savings_lm)
```

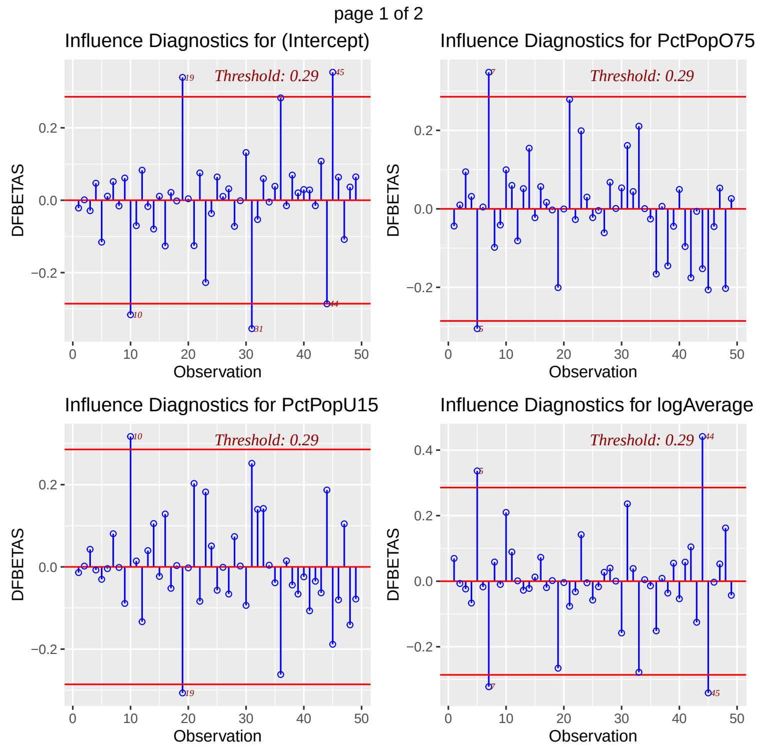


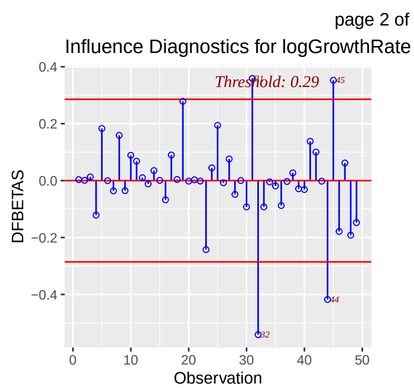



```
# DFFITs plot:
olsrr::ols_plot_dffits(new_savings_lm)
```



```r
# DFBETAs panel:
olsrr::ols_plot_dfbetas(new_savings_lm)
```





Look at final model

Notice that only PopU15 and logGrowthRate had β coefficients that were significant according to the t-test. Thus, we might want to create our final model with only the two significant variables:

```
final_savings_lm <- lm(SavingsRatio ~ PctPopU15 + logGrowthRate, data = new_savings)

summary(final_savings_lm)
```

```
##
## Call:
## lm(formula = SavingsRatio ~ PctPopU15 + logGrowthRate, data = new_savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9342 -2.6413  0.2752  1.8731 10.0690
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.27955    2.40166   5.946 3.49e-07 ***
## PctPopU15     -0.18046    0.05915  -3.051  0.00378 **
## logGrowthRate  1.45209    0.71058   2.044  0.04675 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.69 on 46 degrees of freedom
## Multiple R-squared:  0.2632,Adjusted R-squared:  0.2312
## F-statistic: 8.217 on 2 and 46 DF,  p-value: 0.0008884
```