preliminary_model

Erin Xu & Dora Dong

2025-05-19

Load data and check if needed to clean.

```
## [1] TRUE
```

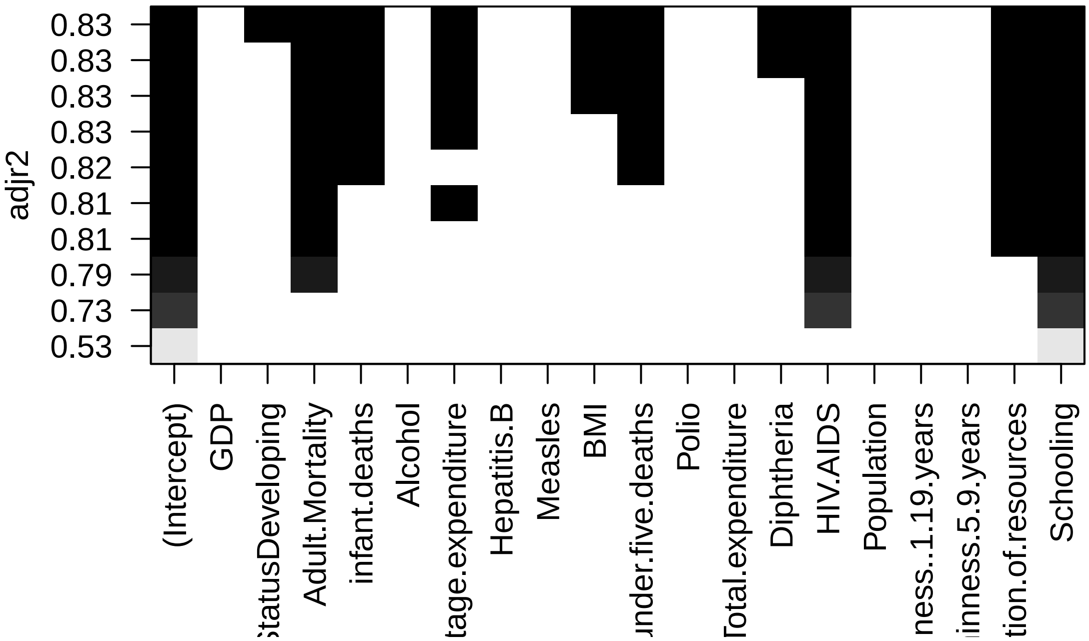
```
## [1] 2563
```

```
##                         Country                            Year 
##                               0                               0 
##                          Status                 Life.expectancy 
##                               0                              10 
##                 Adult.Mortality                   infant.deaths 
##                              10                               0 
##                         Alcohol          percentage.expenditure 
##                             194                               0 
##                     Hepatitis.B                         Measles 
##                             553                               0 
##                             BMI               under.five.deaths 
##                              34                               0 
##                           Polio               Total.expenditure 
##                              19                             226 
##                      Diphtheria                        HIV.AIDS 
##                              19                               0 
##                             GDP                      Population 
##                             448                             652 
##            thinness..1.19.years              thinness.5.9.years 
##                              34                              34 
## Income.composition.of.resources                       Schooling 
##                             167                             163
```

Decide predictors

Best subset selection

```
best_model <- regsubsets(Life.expectancy ~ GDP + Status + Adult.Mortality + infant.deaths + Alcohol +
                           percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths + Po
                           Total.expenditure + Diphtheria + HIV.AIDS + Population +
                           thinness..1.19.years + thinness.5.9.years +
                           Income.composition.of.resources + Schooling,
                         data = all_data, nvmax = 10)

# Summary and plot
best_summary <- summary(best_model)
plot(best_model, scale = "adjr2")
```



We decided on

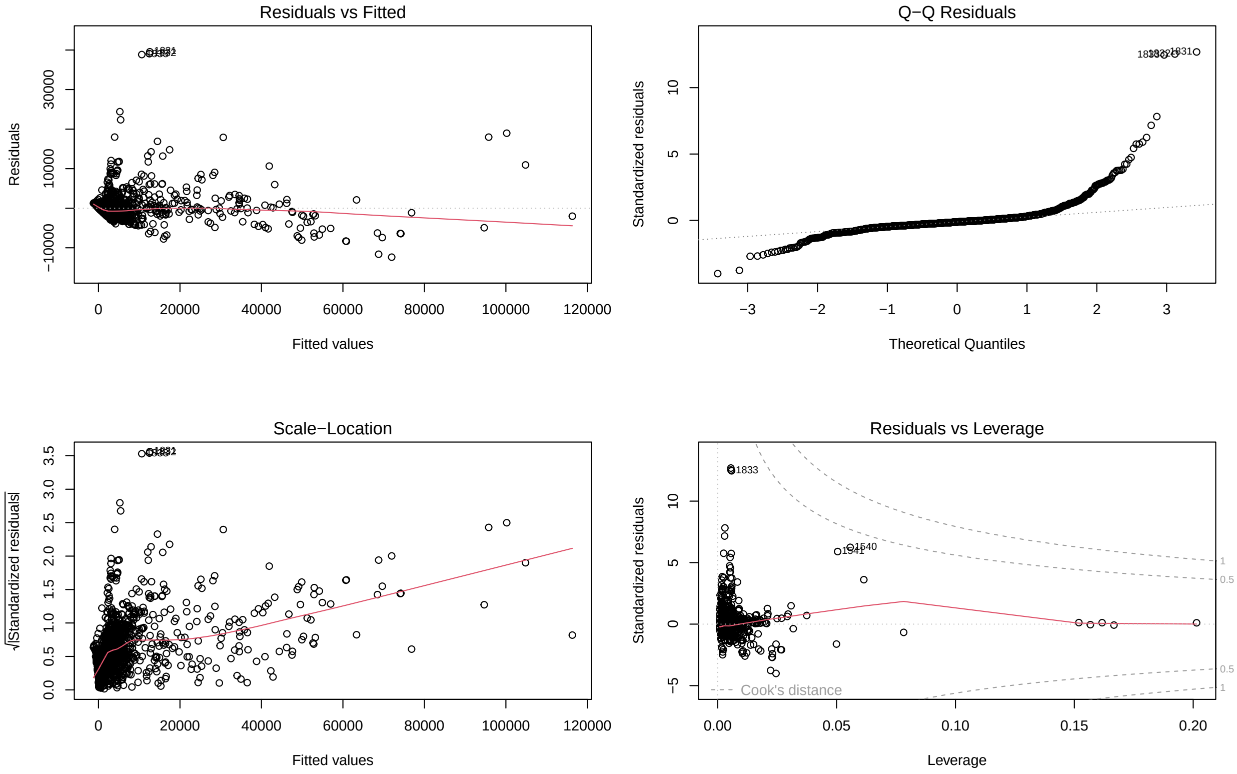
Get the response and predictors.

```
all_data <- read.csv("life_expectancy.csv")
all_data <- na.omit(all_data)
response <- all_data$GDP
x0 <- all_data$percentage.expenditure
x1 <- all_data$Polio
x2 <- all_data$Population
x3 <- all_data$Income.composition.of.resources
x4 <- all_data$Schooling

all_data$Status <- as.factor(all_data$Status)

model <- lm(response ~ Status + x0 + x1 + x2 + x3 + x4, data = all_data)
summary(model)
```

```
##
## Call:
## lm(formula = response ~ Status + x0 + x1 + x2 + x3 + x4, data = all_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12377  -1132   -377    394  39556
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.623e+03  5.743e+02  -2.826 0.004766 **
## StatusDeveloping -7.762e+02  2.680e+02  -2.896 0.003824 **
## x0                5.983e+00  5.115e-02 116.973  < 2e-16 ***
## x1                5.020e+00  3.676e+00   1.366 0.172267
## x2               -3.009e-07  1.095e-06  -0.275 0.783555
## x3                2.082e+03  6.885e+02   3.023 0.002538 **
## x4                1.601e+02  4.713e+01   3.396 0.000699 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3126 on 1642 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9258
## F-statistic:  3428 on 6 and 1642 DF,  p-value: < 2.2e-16
```

Residual plots

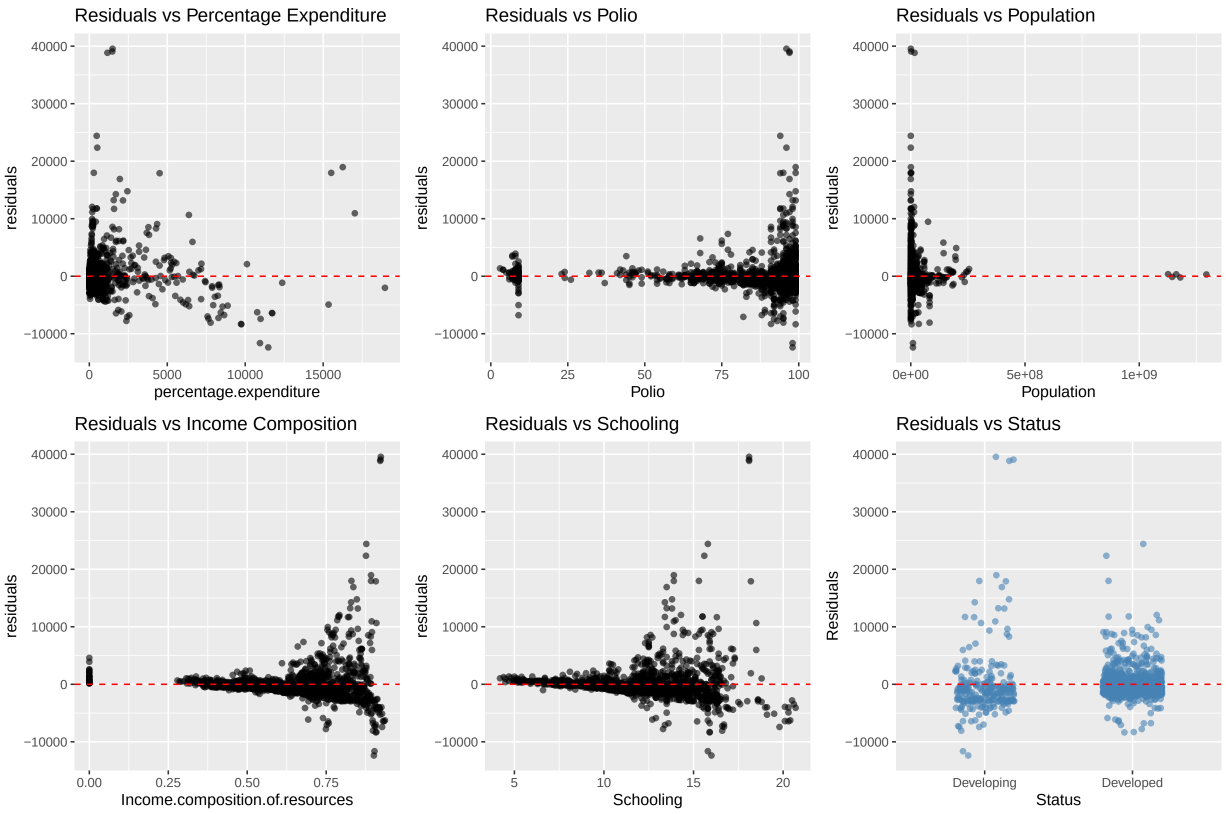
```r
par(mfrow = c(2, 2))
plot(model)
```



# Check for regression assumptions. Residual VS each predictor

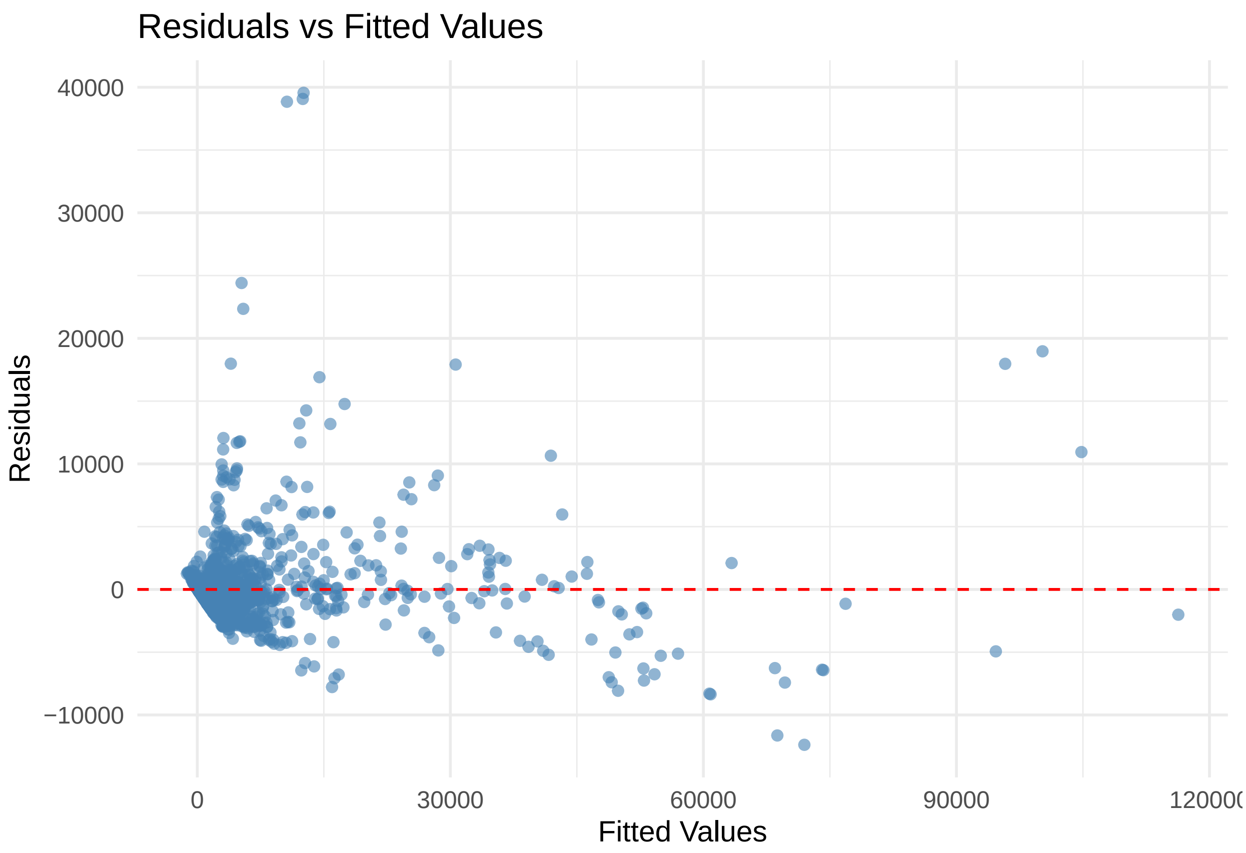
```
# library(ggplot2)
# library(patchwork)

all_data$residuals <- resid(model)
all_data$fitted <- fitted(model)

p1 <- ggplot(all_data, aes(x = `percentage.expenditure`, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Percentage Expenditure")

p2 <- ggplot(all_data, aes(x = Polio, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Polio")

p3 <- ggplot(all_data, aes(x = Population, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Population")

p4 <- ggplot(all_data, aes(x = `Income.composition.of.resources`, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Income Composition")
```

```r
p5 <- ggplot(all_data, aes(x = Schooling, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Schooling")

p6 <- ggplot(all_data, aes(x = as.factor(Status), y = residuals)) +
  geom_jitter(width = 0.2, alpha = 0.6, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  scale_x_discrete(labels = c("Developing", "Developed")) +
  labs(title = "Residuals vs Status",
       x = "Status",
       y = "Residuals")

(p1 | p2 | p3) / (p4 | p5 | p6)
```

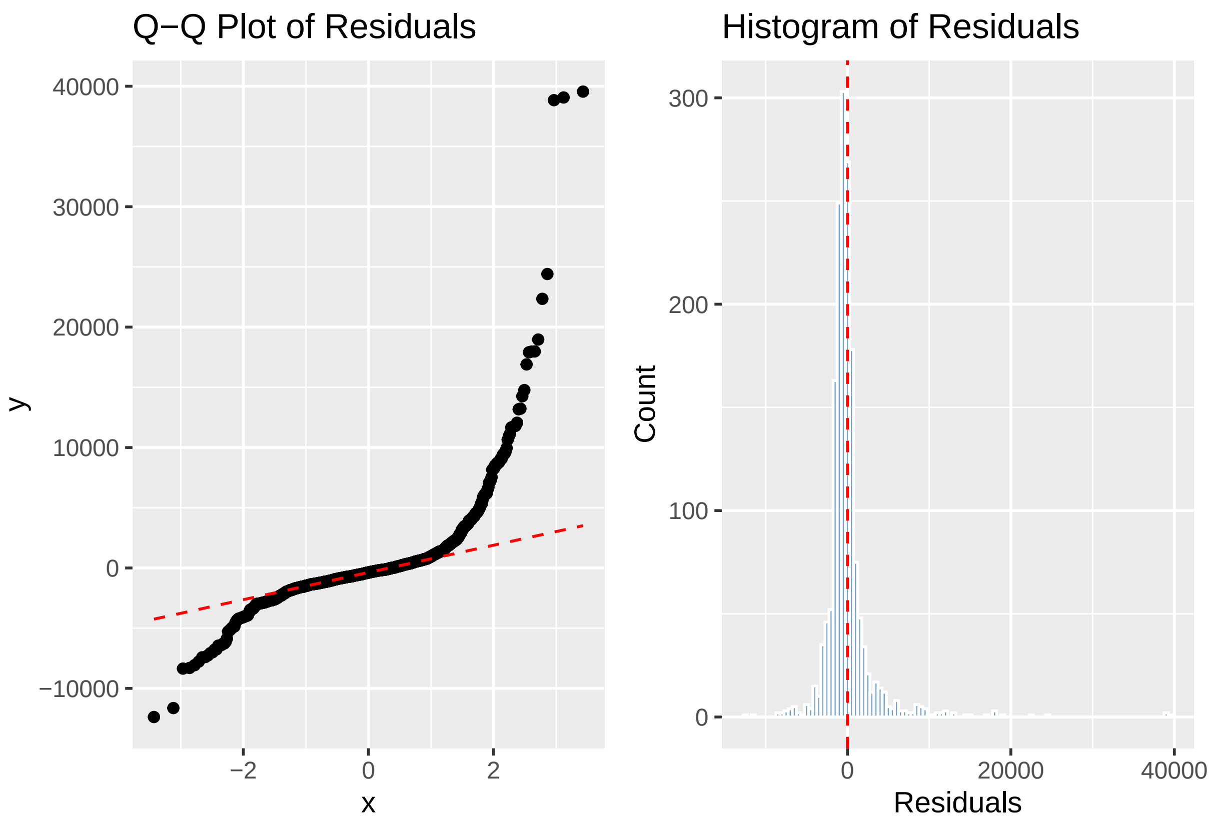


Residual VS Fitted plot

```r
ggplot(all_data, aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()
```



```
qq_plot <- ggplot(all_data, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red", linetype = "dashed") +
  labs(title = "Q-Q Plot of Residuals")

hist_plot <- ggplot(all_data, aes(x = residuals)) +
  geom_histogram(binwidth = 500, fill = "steelblue", color = "white", alpha = 0.7) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Count")

qq_plot | hist_plot
```



# Check for influential points, high leverage points

```
# Set threshold
predictors <- c("Status", "percentage.expenditure", "Polio",
                "Total.expenditure", "Income.composition.of.resources", "Schooling")
cook_thresh <- 4 / nrow(all_data)
lev_thresh <- 2 * mean(all_data$leverage)
```

```
## Warning in mean.default(all_data$leverage): argument is not numeric or logical:
## returning NA
```

```
# Logical vector for flagged rows
flagged_index <- all_data$cooksD > cook_thresh | all_data$leverage > lev_thresh

# Directly subset with logical index
print(all_data[flagged_index, c("Country", predictors)])
```

```
## [1] Country                         Status
## [3] percentage.expenditure          Polio
## [5] Total.expenditure               Income.composition.of.resources
## [7] Schooling
## <0 rows> (or 0-length row.names)
```