Multiple Linear Regression (Model Selection and Model Checking) - Lab

Blake Pappas

December 17, 2023

Savings Rates in 50 Countries

The savings data frame has 50 rows (countries) and 5 columns (variables):

1. : savings rate - personal saving divided by disposable income *This variable will be used as the response*
2. : percent population under age of 15
3. : percent population over age of 75
4. : per-capita disposable income in dollars
5. : percent growth rate of dpi

The data is averaged over the period 1960-1970.

Data Source: Belsley, D., Kuh. E. and Welsch, R. (1980) *Regression Diagnostics* Wiley.

Load the dataset.

Code:

```
data(savings, package = "faraway")
head(savings)
```

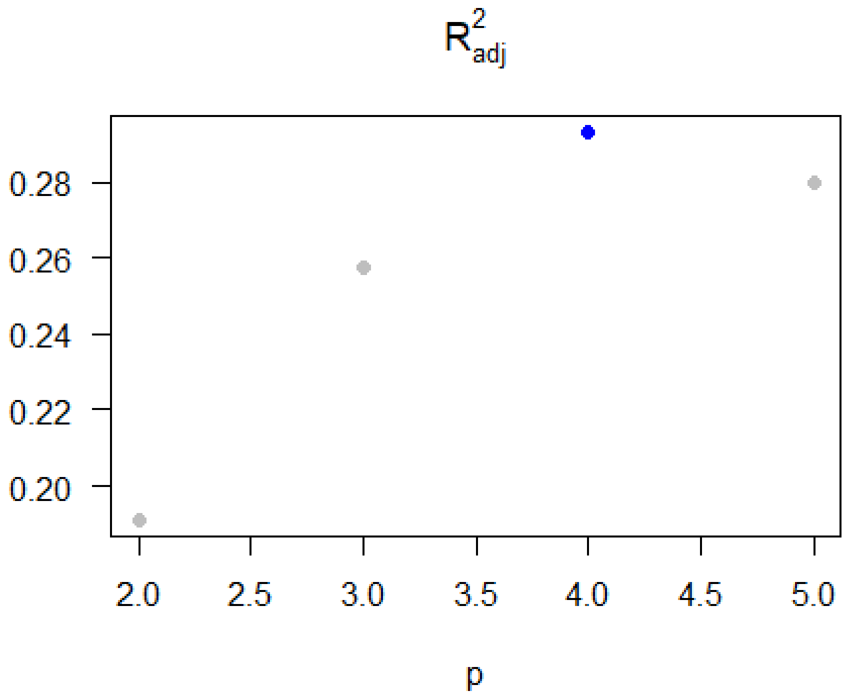
```
##              sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
```

1. Perform the best subset selection and select the "best" model using R^2_{adj}.

Code:

```
library(tidyverse)
library(caret)
library(leaps)
models <- regsubsets(sr ~ ., data = savings) # regsubsets = the function for
```

```r
model selection
summary(models) # Gives best model based on the number of predictors

## Subset selection object
## Call: regsubsets.formula(sr ~ ., data = savings)
## 4 Variables  (and intercept)
##       Forced in Forced out
## pop15     FALSE      FALSE
## pop75     FALSE      FALSE
## dpi       FALSE      FALSE
## ddpi      FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          pop15 pop75 dpi ddpi
## 1  ( 1 ) "*"   " "   " " " "
## 2  ( 1 ) "*"   " "   " " "*"
## 3  ( 1 ) "*"   "*"   " " "*"
## 4  ( 1 ) "*"   "*"   "*" "*"

res.sum <- summary(models)

criteria <- data.frame(
  Adj.R2 = res.sum$adjr2,
  Cp = res.sum$cp,
  BIC = res.sum$bic)

criteria

##      Adj.R2       Cp       BIC
## 1 0.1910048 7.906993 -3.805036
## 2 0.2574811 4.446603 -5.232912
## 3 0.2932620 3.130920 -4.865619
## 4 0.2796525 5.000000 -1.098852

# Plot of Adjusted R-Squared
plot(2:5, criteria$Adj.R2, las = 1, xlab = "p", ylab = "", pch = 16, col =
"gray",
     main = expression(R['adj']^2))
points(4, criteria$Adj.R2[3], col = "blue", pch = 16)
```



Answer: The best model using R^2_{adj} is the third model, which uses the `pop15`, `pop75`, and `ddpi` as the predictors.

2. Perform a stepwise selection using AIC.

Code:

```
full <- lm(sr ~ ., data = savings)
step(full, direction = "both")

## Start:  AIC=138.3
## sr ~ pop15 + pop75 + dpi + ddpi
##
##          Df Sum of Sq    RSS    AIC
## - dpi     1     1.893 652.61 136.45
## <none>                650.71 138.30
## - pop75   1    35.236 685.95 138.94
## - ddpi    1    63.054 713.77 140.93
## - pop15   1   147.012 797.72 146.49
##
## Step:  AIC=136.45
## sr ~ pop15 + pop75 + ddpi
##
##          Df Sum of Sq    RSS    AIC
## <none>                652.61 136.45
## - pop75   1    47.946 700.55 137.99
## + dpi     1     1.893 650.71 138.30
```

```
## - ddpi   1    73.562 726.17 139.79
## - pop15  1   145.789 798.40 144.53
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)
##
## Coefficients:
## (Intercept)        pop15        pop75         ddpi
##     28.1247      -0.4518      -1.8354       0.4278
```

Answer: Using the stepwise selection, we are brought to an AIC of 136.45, which uses `pop15`, `pop75`, and `ddpi` as the predictors for the regression having the best goodness of fit.

3. Perform a general linear F-test (with $\alpha = 0.1$) to choose between the full model (i.e., using the all 4 predictors) and the reduced model that includes `pop15`, `pop75`, and `ddpi` as the predictors.

Code:

```
# Reduced Model
reduce <- lm(sr ~ pop15 + pop75 + ddpi, data = savings)
summary(reduce)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2539 -2.6159 -0.3913  2.3344  9.7070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.1247     7.1838   3.915 0.000297 ***
## pop15        -0.4518     0.1409  -3.206 0.002452 **
## pop75        -1.8354     0.9984  -1.838 0.072473 .
## ddpi          0.4278     0.1879   2.277 0.027478 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.767 on 46 degrees of freedom
## Multiple R-squared:  0.3365, Adjusted R-squared:  0.2933
## F-statistic: 7.778 on 3 and 46 DF,  p-value: 0.0002646
```

```
# Full Model
full <- lm(sr ~ ., data = savings)
summary(full)
```

```
##
## Call:
## lm(formula = sr ~ ., data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

```
# General Linear F-Test
anova(reduce, full)
```

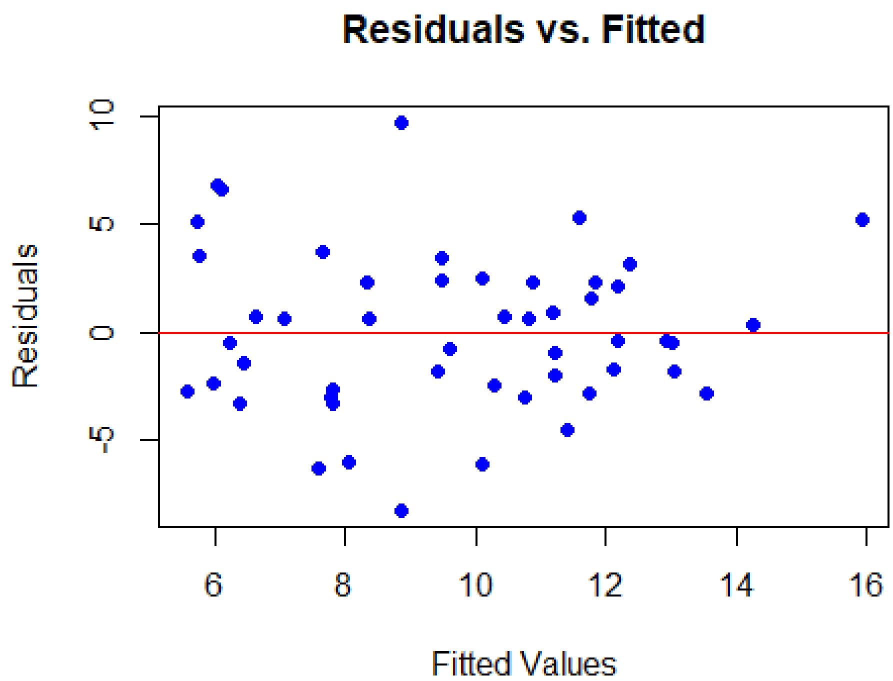
```
## Analysis of Variance Table
##
## Model 1: sr ~ pop15 + pop75 + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     46 652.61
## 2     45 650.71  1    1.8932 0.1309 0.7192
```

Answer: Since the p-value of this general linear F-test is greater than α, we fail to reject H0 and conclude that we do not have sufficient evidence to support that at least one of the three regression coefficients is not equal to 0.

4. Make a residual plot of the model selected by *AIC* and comment on the model assumptions.

Code:

```
mod <- lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)
plot(mod$fitted.values, mod$residuals, pch = 16, col = "blue", xlab = "Fitted
Values", ylab = "Residuals", main = "Residuals vs. Fitted")
abline(h = 0, col = "red")
```

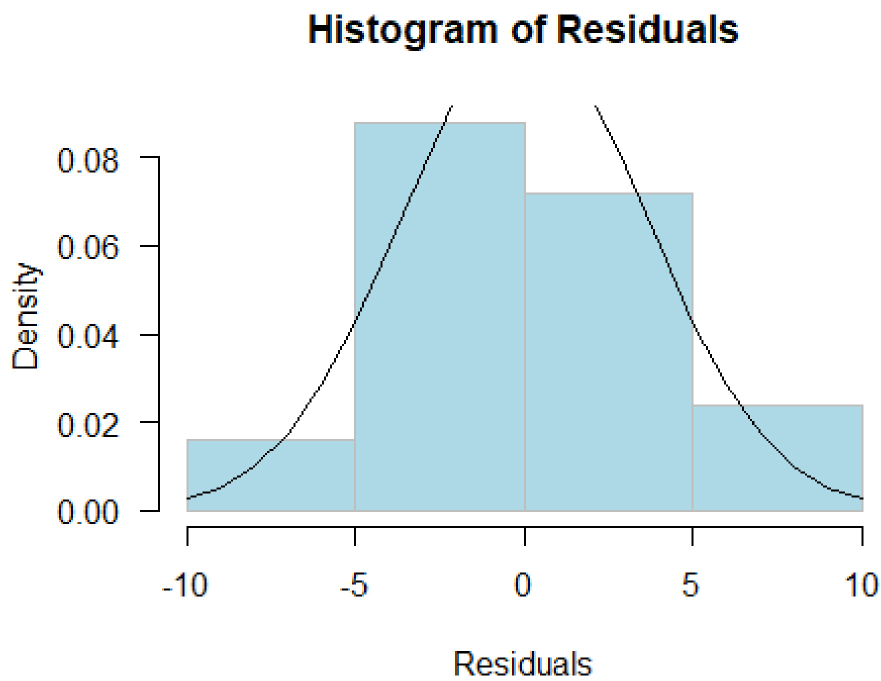


Answer: There is no major concern with the model assumptions, as the plot of the residuals appears to be random.

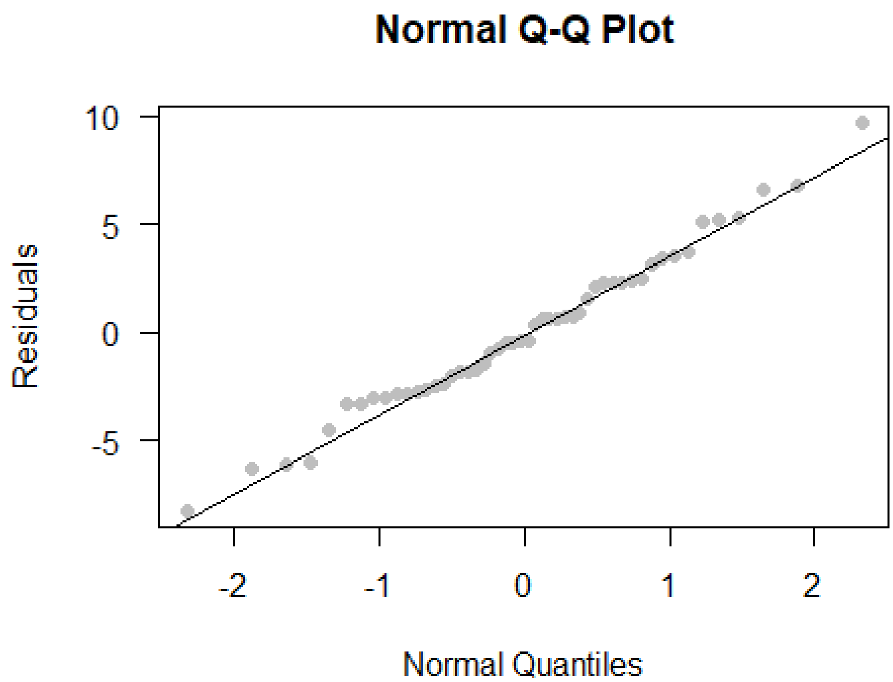
5. Use both a histogram and qqplot to examine the normality assumption on error.

Code:

```
# Histogram
(sd <- sd(mod$residuals))

## [1] 3.649451

par(las = 1)
hist(mod$residuals, 5, prob = T, col = "lightblue", border = "gray", main =
"Histogram of Residuals", xlab = "Residuals")
xg <- seq(-10, 10, 1)
yg <- dnorm(xg, 0, sd)
lines(xg, yg)
```




```r
# qqplot
qqnorm(mod$residuals, pch = 16, las = 1, col = "gray", xlab = "Normal
Quantiles", ylab = "Residuals")
qqline(mod$residuals)
```

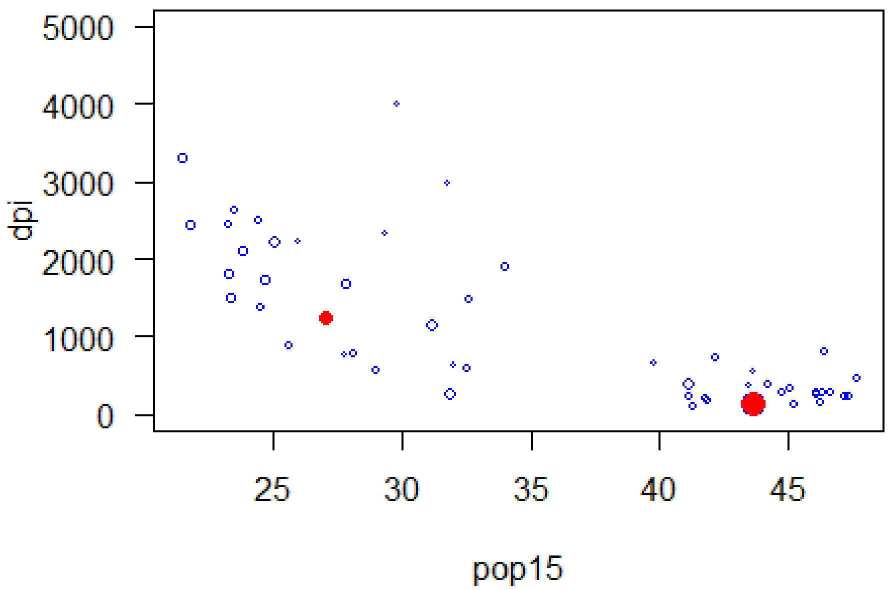


Answer: There is no major concern regarding normality with this model. The distribution of the residuals appear to be approximately normally distributed. The Normal Q-Q Plot appears to run closely (in an S-shaped pattern) to the trend line,with little deviation.

6. Calculate the leverage values to check if there is any high leverage points (i.e., $h > \frac{2p}{n}$).

Code:

```
step_savings <- step(full, trace = F) # Trace = full
X <- model.matrix(step_savings) # Model Design Matrix
H <- X %*% solve((t(X) %*% X)) %*% t(X)
lev <- hat(X) # Calculates leverage
high_lev <- which(lev >= 2 * 3 / 30) # Finds the high leverage values
attach(savings)

# Plot of Leverage Points
par(las = 1)
plot(pop15, dpi, cex = sqrt(5 * lev), col = "blue", ylim = c(0, 5000))
points(pop15[high_lev], dpi[high_lev], col = "red", pch = 16,
       cex = sqrt(5 *lev[high_lev]))
```

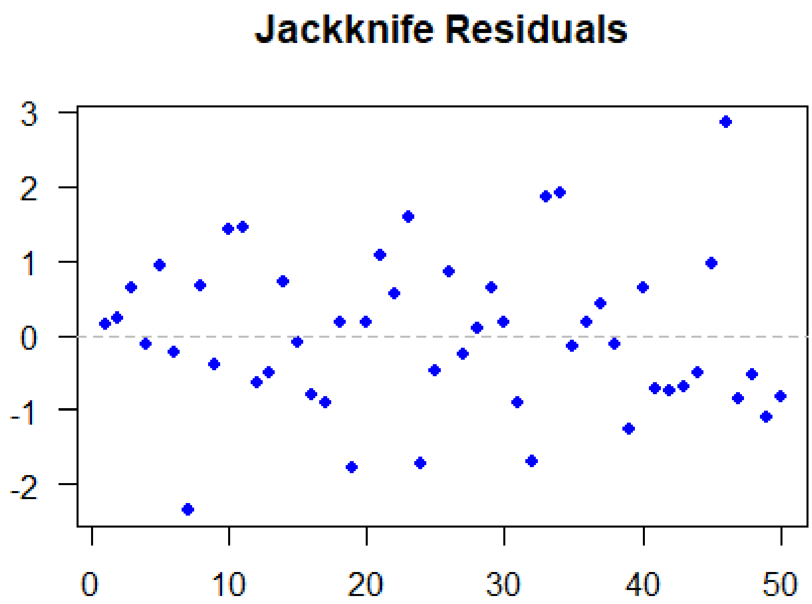


Answer: There are two high leverage points between `pop15` and `dpi`. They exist in observations 23 an 49.

7. Compute jackknife residuals to identify outlier(s).

Code:

```
jack <- rstudent(step_savings)

par(las = 1)
plot(jack, pch = 16, cex = 0.8, col = "blue", main = "Jackknife Residuals",
     xlab = "", ylab = "")
abline(h = 0, lty = 2, col = "gray")
```

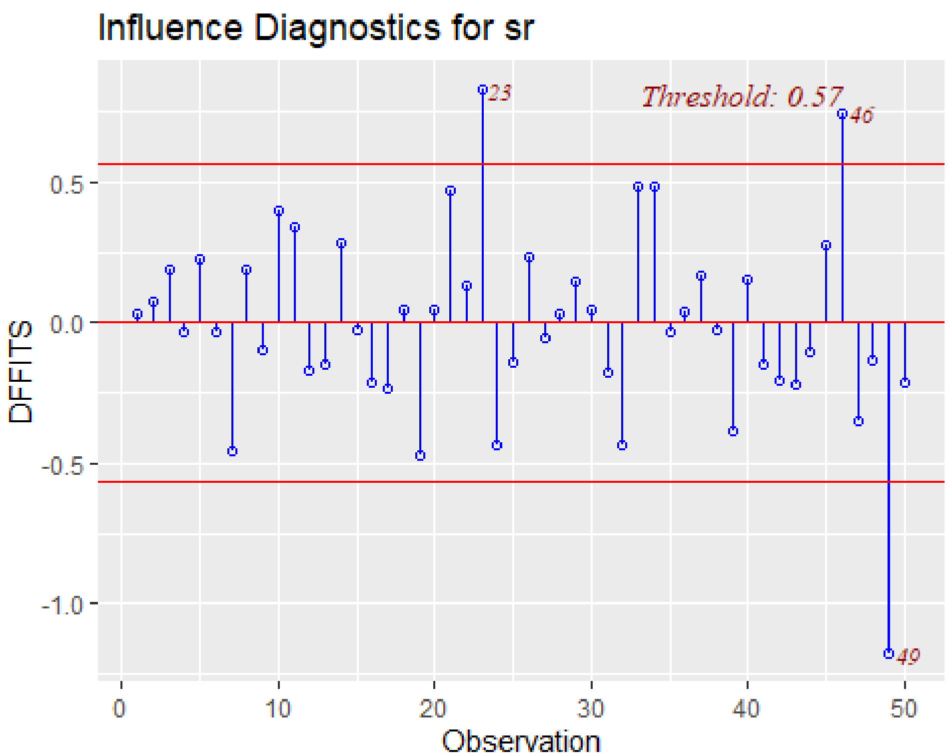


Answer: Looking at the graph, there do not appear to be many outliers.

8. Identifying influential observations by computing DFFITS.

Code:

```
library(olsrr)
ols_plot_dffits(step_savings)
```



Answer: There are three influential observations for sr: Observations 23, 46, and 49.