36-402 Homework 8

James “Morgan” Hawkins

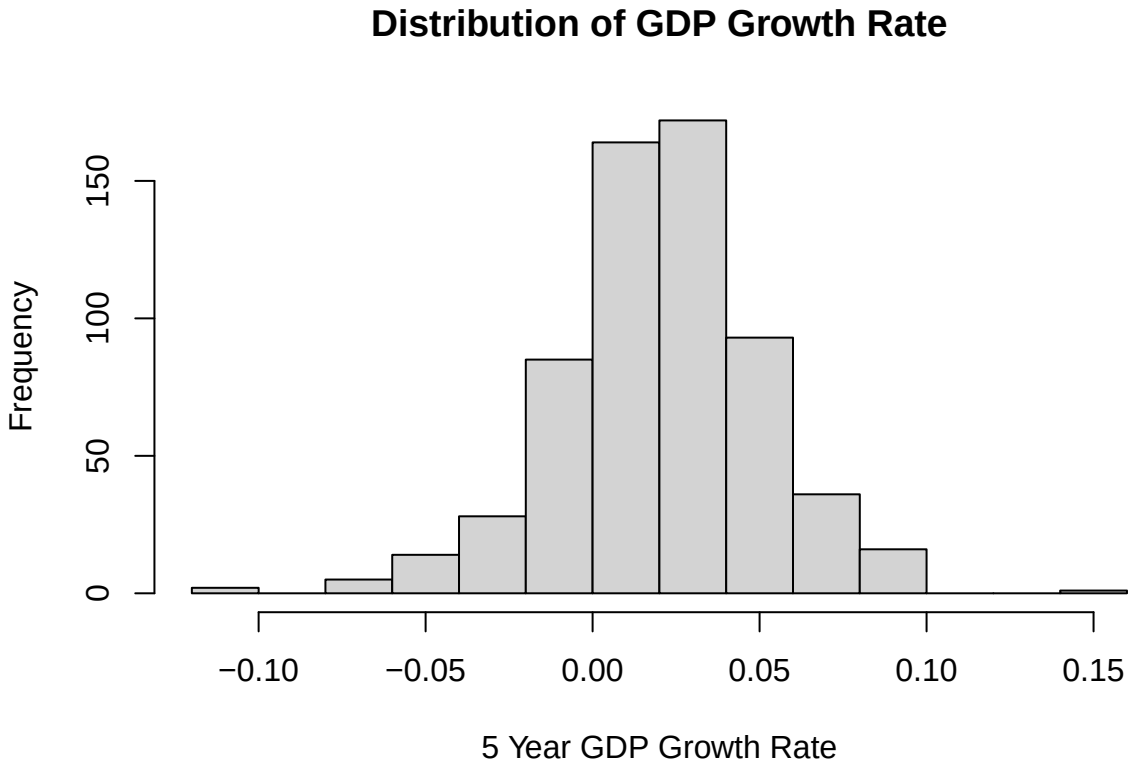
Friday April 7, 2023

Problem 1

```
data(oecdpanel)
gdp = oecdpanel
```

Problem 1 (a)

```
hist(gdp$growth, main = "Distribution of GDP Growth Rate",
     xlab = "5 Year GDP Growth Rate")
```



```
#sd(gdp$growth)
```

In the plot above we see that our response variable growth appears to have a unimodal, symmetrical distribution. Growth has a sample mean of .020 and sample standard deviation of .030 with a few outliers at around 15% and -10%.

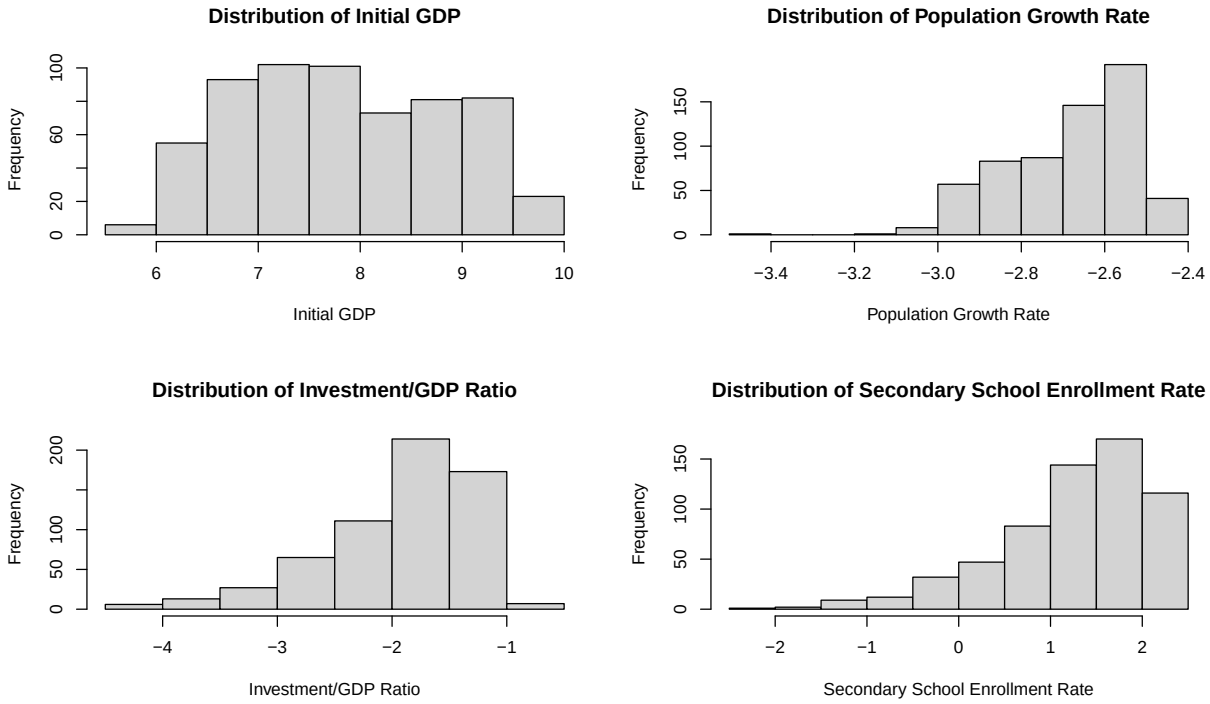
```
par(mfrow = c(2,2))

var.des = c("Initial GDP",
            "Population Growth Rate",
            "Investment/GDP Ratio",
            "Secondary School Enrollment Rate")



for(col in 4:7){
  #cat(colnames(gdp)[col],mean(as.numeric(gdp[,col])), sd(as.numeric(gdp[,col])),"\n\n")
  hist(as.numeric(gdp[,col]), main = paste("Distribution of",var.des[col-3]), xlab = var.des[col-3])

}
```



In the plot above we see that the distribution of Initial GDP has a somewhat unimodal distribution that is slightly left-skewed. Initial GDP has a mean of 7.82 and a standard deviation of 1.01 with no obvious outliers. Population growth rate has a unimodal left-skewed distribution. Population growth rate has a mean of -2.68 and a standard deviation of .152 with an outlier at -3.4. Investment/GDP ratio has a unimodal left-skewed distribution. Investment/GDP ratio has a mean of -1.93 and a standard deviation of .656 with no clear outliers. Secondary school enrollment rate has a unimodal left-skewed distribution. Secondary school enrollment rate has a mean of 1.26 and a standard deviation of .825 with no clear outliers.

```
table(gdp$oecd)/(dim(gdp)[1])
```

```
##
##         0         1
## 0.7386364 0.2613636
```

```
table(gdp$year)/(dim(gdp)[1])
```

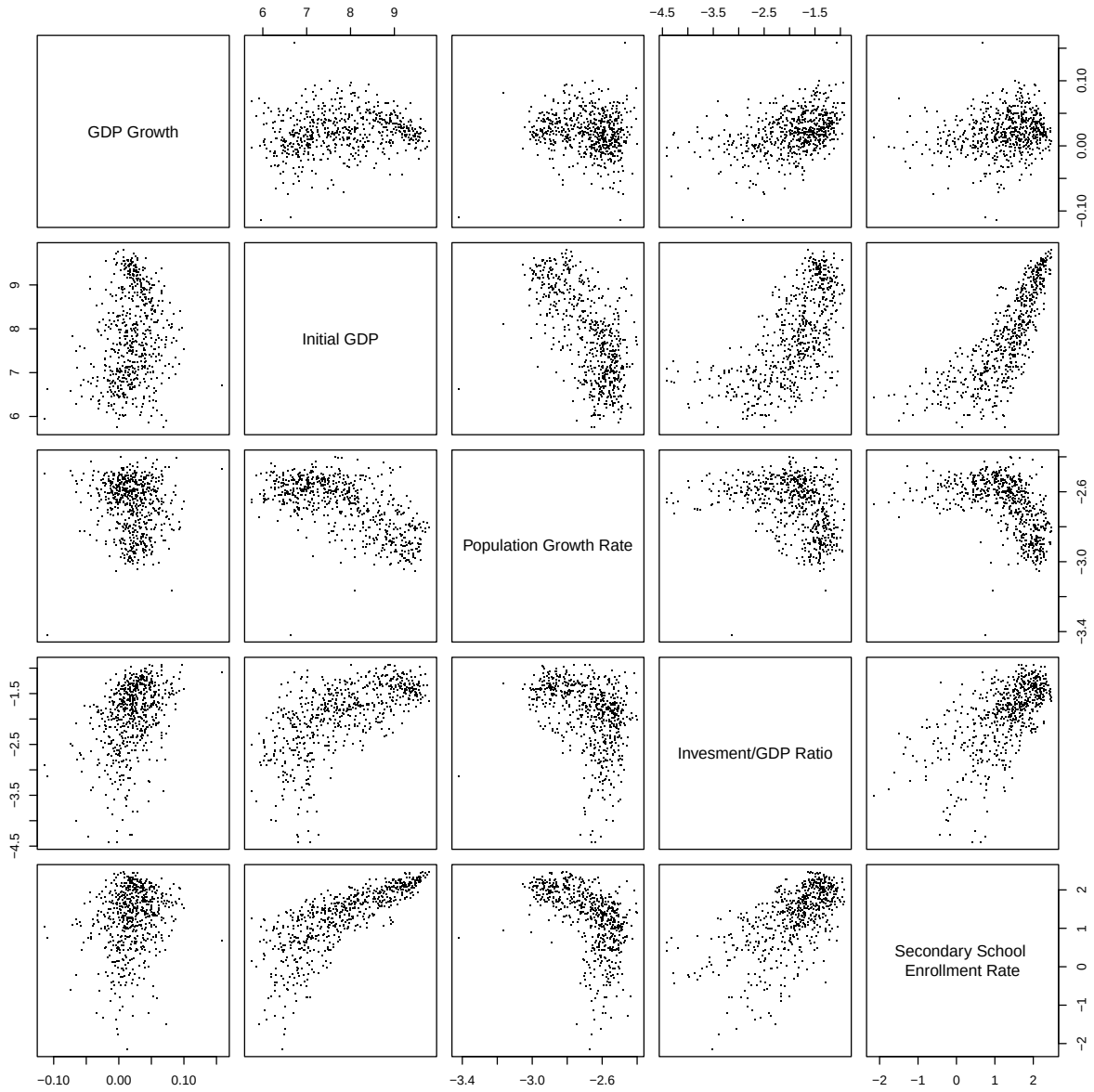
```
##
##      1965      1970      1975      1980      1985      1990      1995
## 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571
```

```
is.na(gdp) %>% sum
```

```
## [1] 0
```

In our data around 26.1% of our cases are countries that were a member of OECD in the time frame of the case. We also see that year is uniformly distributed and all countries in the data set have data for each year. There are no missing data points.

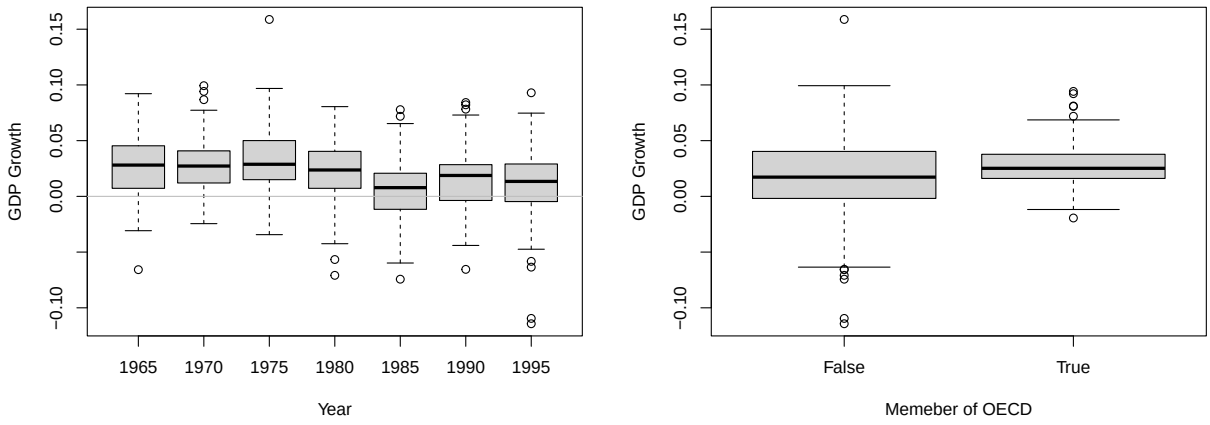
```
c("GDP Growth", "Initial GDP", "Population Growth Rate", "Invesment/GDP Ratio", "Secondary School \nEnr
  pairs(gdp[,-c(2,3)], pch = '.', labels = .)
```



In the plot of continuous variables pairs above we notice the following possible relationships. A negative linear relationship between initial GDP and population growth rate, positive relationship between investment/GDP ratio and GDP growth, positive non-linear relationship between initial GDP and investment GDP ratio, negative non-linear relationship between population growth rate and secondary school enrollment rate, positive relationship between secondary school enrollment rate and investment/GDP ratio, and a positive non-linear relationship between initial GDP and secondary school enrollment rate.

```
par(mfrow = c(1,2))

boxplot(growth ~ year, data = gdp, ylab = "GDP Growth", xlab = "Year")
abline(0,0, col = 'grey')
boxplot(growth ~ oecd, data = gdp, ylab = "GDP Growth", xlab = "Memeber of OECD", xaxt = 'n')
axis(1, at = c(1,2), labels = c("False", "True"))
```



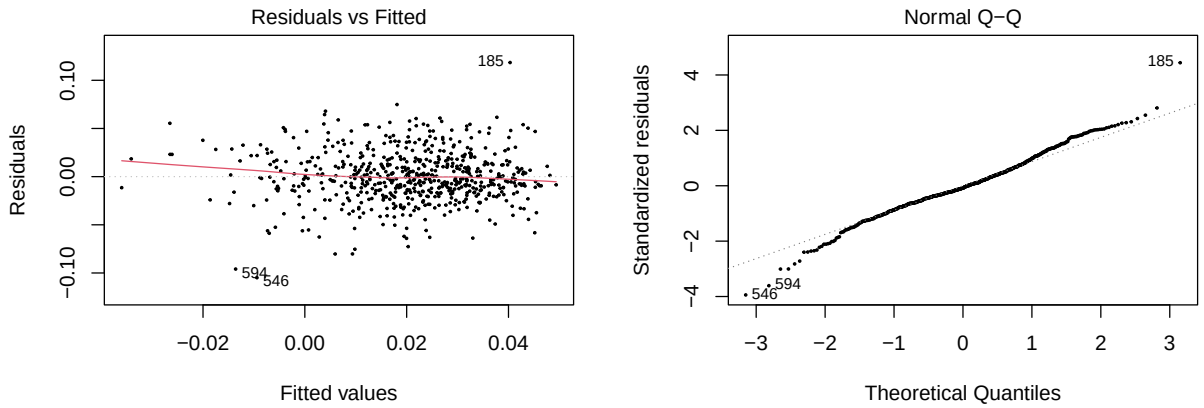
Above we see that all 7 time periods have average GDP growth rates above 0 with 1970-1974 having the highest average growth rate among countries while 1980-1984 had the lowest average growth rate among countries. We also see that countries that are in the OECD have higher average growth rates as well as less variance in their 5 year growth rates. We also notice that countries in the OECD have more outliers on the right side of the distribution of GDP growth while countries not in the OECD have more outliers on the left-tail of the distribution of GDP growth.

Problem 1 (b)

```
model.1 = lm(growth ~ year + inv, data = gdp)
model.1 %>% summary
```

```
##
## Call:
## lm(formula = growth ~ year + inv, data = gdp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.104901 -0.015860 -0.002044  0.015584  0.118443
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.5662800  0.2132593   7.344 6.61e-13 ***
## year        -0.0007624  0.0001077  -7.080 3.97e-12 ***
## inv          0.0188115  0.0016432  11.448  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02673 on 613 degrees of freedom
## Multiple R-squared:  0.2275, Adjusted R-squared:  0.225
## F-statistic: 90.27 on 2 and 613 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,2))
plot(model.1,1:2, pch = 19, cex = .25)
```



On the residual vs. fitted plot we see that the residuals appear to have constant variance and mean 0. On the normal QQ plot we see that our residuals are also approximately normal. However, the empirical distribution appears to have slightly wider tails than the assumed normal distribution, especially on the left tail.

```
par(mfrow = c(1,2))
boxplot(residuals(model.1) ~ gdp$year,
        xlab = "End Year", ylab = "Residual", main = "Residuals vs Time Period")
abline(0, 0, col = 'grey')
plot(gdp$inv,residuals(model.1), pch = 19, cex = .25,
     xlab = "Investment/GDP ratio", ylab = "Residual", main = "Investment / GDP Ratio vs Time Period")
abline(0, 0, col = 'grey')
```