Colin White
March 12, 2021

HOMEWORK ANALYSIS #6

What spurs economic growth? These data consist of economic data from 88 countries The dataset gdp_sub.csv contains the following information for each country:

Variable Name	Description
y	Growth of GDP per capita at purchasing power parities between 1960 and 1996.
everything else	See table under content tab (48 indicators)

The predictors are described in more detail in variable table.pdf. In this analysis, we treat GDP Growth per capita at purchasing power parities (y in the dataset) between 1960 and 1996 is the response variable. The other 48 variables are economic variables that we use as explanatory variable with the goal of explaining GDP growth. Many of the covariates are binary (or dummy variables) but most are quantitative. Our goal is to understand what relationships there are between economic indicators and gdp growth.

1. In your own words, summarize the overarching problem and any specific questions that need to be answered using the GDP dataset. Discuss how statistical modeling will be able to answer the posed questions.

We want to spur economic growth. In order to do that we want to find a way to predict economic growth to know it's causes. We are trying to see if various variables from (gdp_sub.csv) can predict economic growth.

If the gdp_sub.csv data passes the LINE assumptions, we can fit a MLR model to the data and use it to predict the economic growth of countries. This would work by taking a measure of one or more variables (economic indicators) from gdp_sub.csv, plugging it into the MLR model, and the model predicting the economic growth of that country.

2. Calculate variance inflation factors. What proportion of the 48 variables pose an issue? Comment on what affect collinearity would have on a regression analysis.

```
> car::vif(lm.model)
   ABSLATIT     AIRDIST       AVELF      BUDDHA       CIV72      COLONY      CONFUC
  15.029613    5.683950    7.592927    6.471788    7.169144    9.180599    3.600737
     DENS60     DENS65C     DENS65I    DPOP6090        EAST       ECORG      EUROPE
   3.231013   11.708255    3.296389   25.660756    8.193883    2.275256   41.795108
   FERTLDC1        GDE1    GDPCH60L     GEEREC1     GOVSH61       GVR61         H60
  52.681370  154.956775   14.877428   24.425185 1232.085437 1354.124586    6.825100
     HERF00     IPRICE1        LAAM    LANDLOCK     LIFE060      LT100CR    MALFAL66
   2.780911    2.364438   30.069784    3.718136   24.680309    7.684500    9.045881
   NEWSTATE    OPENDEC1    OTHFRAC         P60     PRIGHTS     POP1560       POP60
   9.009757   11.543049    4.102418    8.184369   11.216429   47.124362    3.260354
    POP6560    PRIEXP70        RERD     REVCOUP     SAFRICA      SIZE60       SPAIN
  26.551867    7.606801    4.448601    2.855710   15.076620   18.964158   11.960740
   TROPICAR     TROPPOP     WARTIME     WARTORN     YRSOPEN    ZTROPICS
  21.126715   14.056126    3.632901    3.400395   10.112859    3.941023
```

About half (24 of the 48) variables pose an issue. This is because they're variance inflation factors are near or above 10. This means that if all variable were included, the model would be over fitted

3. Intuitively discuss the challenge of variable selection a model with 48 predictors using only 88 observations. Which approaches could you feasibly use for variable selection?

In this situation there is a challenge of variable selection because with 48 variables, 88 is an especially low number of observations. With this few observations, it will be more challenging to determine collinearity and each variable associating with the GDP.

48 variable is too many for using "Best" because it would have been too computationally expensive. It would have been 2^48 permutations. I will therefore be using "Forward" in order not to overwhelm my computers processing capabilities.

4. Use a variable selection (not exhaustive) technique to determine a MLR model that will answer the questions posed in #1 (do NOT consider any interactions). Justify your choice of a model comparison criterion (e.g. state why you chose to base your variable selection procedure on AIC vs. BIC).

I used AIC because the goal is prediction, and not BIC because BIC is mainly used if the goal is inference.

5. Write out (in mathematical form with greek letters) your selected MLR model from #4, discussing clearly what variable are in your model. Clearly state any assumptions you are using in your model. Provide an interpretation of the regression coefficient for one quantitative and one binary/dummy variable term included in your model. Discuss how your model, after fitting it to the data, will be able to answer the questions in this problem.

For any data to be suitable to analyze with a MLR model, it needs to pass 4 assumptions. These assumptions are that there is a linear relationship between covariates (economic indicators) and response (GDP), that errors are independent of one another and of the covariates (economic indicators), that errors are normally distributed, and that there is equal variance for all ϵi.

$Y = ß_0 + ß_1*x_1 + ß_2*x_2 + ß_3*x_3 + ß_4*x_4 + ß_5*x_5 + ß_6*x_6 + ß_7*x_7 + ß_8*x_8 + ß_9*x_9 + ß_{10}*x_{10} + ß_{11}*x_{11} + ß_{12}*x_{12} + ß_{13}*x_{13} + ß_{14}*x_{14} + ß_{15}*x_{15} + ß_{16}*x_{16} + ß_{17}*x_{17} + \epsilon_i$ where $\epsilon_i \sim {}^{iid}N(0, \sigma^2)$

Y = Predicted GDP
$ß_0$ = The predicted GDP on average if all the measurements were at zero. It is the Y axis intercept.

$ß_1$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_1 ((snowfall) in inches)
x_1 = CONFUC = Fraction of population Confucian

$ß_2$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_2
x_2 = DENS60 = Population per area in 1960.

ß$_3$ = Holding all other x's constant this is the amount the predicted GDP increase on average if the country is categorized as an East Asian countries (X_3).
x_3 = East = East Asian countries (BINARY)

ß$_4$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_4
x_4 = GDPCH60L = Logarithm of GDP per capita in 1960.

ß$_5$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_5
x_5 = GEEREC1 = Average share public expenditures on education as fraction of GDP between 1960 and
1965.

ß$_6$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_6
x_6 = GVR61 = Share of expenditures on government consumption to GDP in 1961.

ß$_7$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_7
x_7 = H60 = Enrollment rates in higher education.

ß$_8$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_8
x_8 = IPRICE1 = Average investment price level between 1960 and 1964 on purchasing power parity basis

ß$_9$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_9
x_9 = LIFE060 = Life expectancy in 1960.

ß$_{10}$ = Holding all other x's constant this is the holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{10}
x_{10} = MALFAL66 = Index of malaria prevalence in 1966

ß$_{11}$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{11}
x_{11} = OTHFRAC = Fraction of population speaking foreign language.

ß$_{12}$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{12}
x_{12} = P60 = Enrollment rate in primary education in 1960.

ß$_{13}$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{13}

x_{13} = POP60 = Population in 1960

$ß_{14}$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{13}
x_{14} = RERD = Real exchange rate distortions

$ß_{15}$ = Holding all other x's constant this is the amount the predicted GDP increase on average if the country is categorized as former Spanish colonies (X_{15}).
x_{15} = SPAIN BIANARY = Dummy variable for former Spanish colonies

$ß_{16}$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{13}
x_{16} = TROPICAR = Proportion of country's land area within geographical tropics.

$ß_{17}$ = Holding all other x's constant this is the amount the predicted GDP increase on average per unit increase x_{13}
x_{17} = YRSOPEN = Number of years economy has been open between 1950 and 1994.

ϵ = average of residuals, average distance to mean about line

I would use this fitted model by inputting the economic indicators and interpreting the fitted value output for predicted GDP. This can help advise countries on how to spur economic growth.

6. Fit your model in #5 to the GDP data and summarize the results by displaying the estimated coefficients in a table with 95% confidence intervals for each parameter. Provide an example of how to interpret one 95% confidence intervals for a quantitative explanatory variable and one 95% confidence intervals for a categorical explanatory variable correctly in the context of the problem

Estimated Coefficients Confidence interval table:

		2.5 %	97.5 %
Intercept	$ß_0$ = 5.324e+00	2.2148018	8.433087
CONFUC	$ß_1$ = 3.109e+00	0.0573368	6.1604030
DENS60	$ß_2$ = 1.585e-03	0.0005218	0.0026481
EAST BINARY	$ß_3$ = 1.670e+00	0.7940814	2.5456665
GDPCH60L	$ß_4$ = -9.312e-01	-1.417227	-0.4451906
GEEREC1	$ß_5$ = 2.061e+01	-3.9064192	45.1321210
GVR61	$ß_6$ = -3.839e+00	-7.1729210	-0.5046113
H60	$ß_7$ -7.826e+00	-14.6049779	-1.0464816
IPRICE1	$ß_8$ = -6.818e-03	-0.0112129	-0.0024238
LIFE060	$ß_9$ = 7.149e-02	0.0249661	0.1180178
MALFAL66	$ß_{10}$ = -2.453e-01	-1.1829933	0.6924133
OTHFRAC	$ß_{11}$ = 7.497e-01	0.1506771	1.3487243
P60	$ß_{12}$ = 1.156e+00	-0.170450	2.4831232
POP60	$ß_{13}$ = 4.926e-06	0.0000007	0.0000092

RERD	$\beta_{14} = -7.158e-03$	-0.0133695	-0.0009459
SPAIN (BINARY)	$\beta_{15} = -7.602e-01$	-1.4869131	-0.0335312
TROPICAR	$\beta_{16} = -5.400e-01$	-1.2642749	0.184370
YRSOPEN		-0.3485134	1.4660814

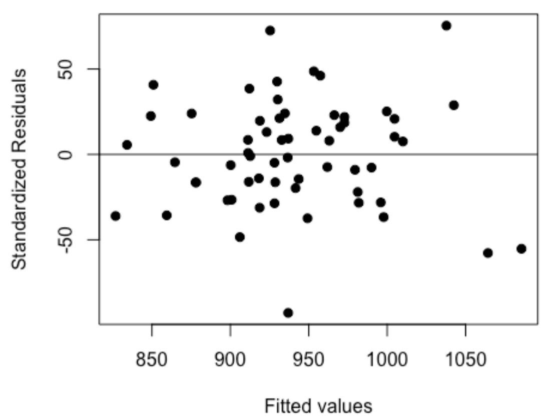
Holding all other x's constant we are 95% confident that GDP will rise between 0.0005218 and 0.002648 on average per unit increase in DENS60 (Population per area in 1960.)

Holding all other x's constant we are 95% confident that GDP will rise between 0.7940814 and 2.5456665 on average if the country is categorized as an East Asian countries (X_3).

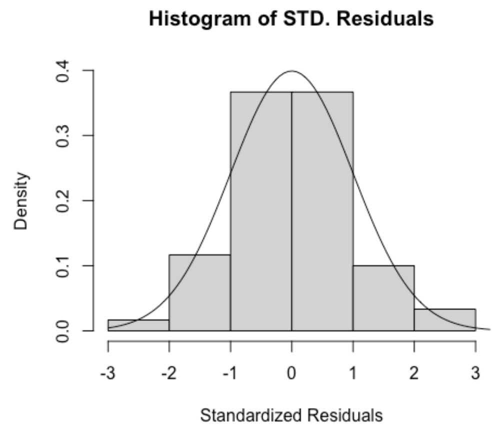
7. Discuss your model assumptions using appropriate graphics or summary statistics. If using graphics, use only a few representative plots.

For any data to be suitable to analyze with a MLR model, it needs to pass 4 assumptions. These assumptions are that there is a linear relationship between covariates and response (GDP), that errors are independent of one another and of the covariates, that errors are normally distributed, and that there is equal variance for all ϵi.

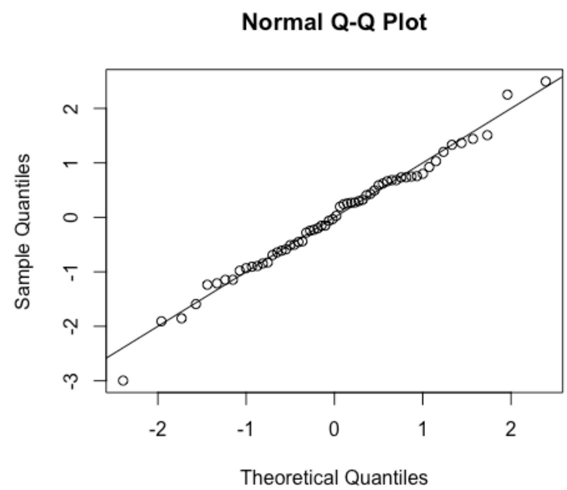
The residuals vs. fitted values scatterplot bellow suggests that the data passes the Linear, Independent and Equal Variance Assumptions.



The histogram of standardized residuals below suggests that the data passes the normality assumption. I propose this is because of the relatively small number of observations, and if more observations were made that histogram would normalize.



The normal quantile-quantile (QQ) plot below also suggests that the data meets the normality assumption.



data: my.best.lm
BP = 5.4602, df = 5, p-value = 0.3623

8. Assess the fit/performance of your model using R2.
The adjusted R-squared value for this model is 0.7814. This means that with this model we can explain 78.14% of variation in GDP. I used adjusted R squared because I variable selected the parameters for this model, and wanted to account for including several variable's

Appendix:

```
#This code is derived from examples in class
library(car)     ## avPlots, vif
library(MuMIn)   ##
install.packages("foreach")
library(bestglm) ##You can also use regsubsets() in library(leaps) but I like this one
```

```
library(lmtest)   ## bptest
library(MASS)     ## stdres
library(car)      ## added-variable plots
library(normtest)
library(knitr)


##########################################
## Read in EI Data and Make Transforms ##
setwd("~/Desktop/1A School/1A Winter 2021/STAT330/HW6")
ei = read.csv("gdp_sub_2.csv", header=TRUE)
head(ei)
pairs(ei[,c(1:)])

##################
## Look at VIFs ##
##################
round(cor(log.ei[,c(16,7:15)]),2)

##### ##### ##### ##### ##### ##### ##### ##### ##### ##### #####
##### VIF(x_j) = 1/(1 - R^2 for covariate j using all other covariates)
##### VIF(\hat{\beta}_j) = 1/(1 - R^2 for covariate j using all other covariates)
##### ##### ##### ##### ##### ##### ##### ##### ##### ##### #####

lm.model = lm(y~.,data=ei)
car::vif(lm.model)
#kable(car::vif(lm.model))

with(log.ei,plot(log.Hydro,log.Nit,pch=19))


X = model.matrix(lm.model)
y = log.ei$AAMort
Xy = cbind(X,y)

var.selection = bestglm(log.ei,IC="BIC",method="exhaustive",
                    TopModels=10) # Y MUST BE THE LAST COLUMN!!!!
my.best.lm = var.selection$BestModel

plot(var.selection$Subsets$BIC,type="b",pch=19,xlab="# of Vars",ylab="BIC")
summary(var.selection$BestModel)
confint(var.selection$BestModel)

var.selection2 = bestglm(log.ei,IC="BIC",method="forward",TopModels=10)
plot(var.selection2$Subsets$BIC,type="b",pch=19,xlab="# of Vars",ylab="BIC")
summary(var.selection$BestModel)
```

```
summary(var.selection2$BestModel)

var.selection$BestModels$Criterion[1]
var.selection2$BestModels$Criterion[1]

############################################
## Perform Forward Selection using AIC  ##
############################################

var.selection = bestglm(ei,IC="AIC",
                    method="forward",TopModels=10)
plot(var.selection$Subsets$AIC,type="b",
    pch=19,xlab="# of Vars",ylab="AIC")
summary(var.selection$BestModel)

########################################################
## Perform Backward Variable Selection using PMSE ##
########################################################

var.selection = bestglm(log.ei,IC="CV",method="exhaustive",
                    TopModels=10,t=1000)

plot(0:15,var.selection$Subsets$CV,type="b",pch=19,xlab="# of Vars",ylab="CV")


# pdf("plot_whatever.pdf",width = 1)
# plot(0:15,var.selection$Subsets$CV,type="b",pch=19,xlab="# of Vars",ylab="CV")
# dev.off()

summary(var.selection$BestModel)

var.selection$BestModel

####################### LOOCV

var.selection.loocv = bestglm(log.ei,IC="LOOCV",method="exhaustive",TopModels=10)

summary(var.selection.loocv$BestModel)

var.selection$BestModel

##each of these should be linear
avPlots(my.best.lm,pch = 20,cex = 0.8)

plot(my.best.lm$fitted.values,my.best.lm$residuals,pch=19, ylab="Standardized Residuals",
xlab="Fitted values")
```

```
abline(a=0,b=0)

## equal variance

plot(my.best.lm$fitted.values,my.best.lm$residuals,pch=19)
abline(a=0,b=0)

plot(ei$CONFUC,my.best.lm$residuals,pch=19)
abline(a=0,b=0,col = "red")

bptest(my.best.lm)

## Normality

hist(stdres(my.best.lm),freq = FALSE, xlab="Standardized Residuals", ylab="Density",
ylim=c(0,.4), main="Histogram of STD. Residuals")
curve(dnorm,from = -3,to = 4,add = TRUE)

qqnorm(stdres(my.best.lm))
abline(0,1)


jb.norm.test(stdres(my.best.lm),nrepl = 1e5)
ks.test(stdres(my.best.lm),"pnorm")

kable(confint(my.best.lm))
```