1. For a long time Macroeconomists have linked life expectancy with the economic well being of a country. In order to test this hypothesis the economic well-being of a country measured in Per Person GDP (PPGDP) was recorded, along with the countries name and whether that is a member of OECD, or part of Africa or neither. These were used and compared to the life expectancy in each country to determine whether or not there is a relationship. Using statistical modeling we will be able to relationship between the effects of country, group, and PPGDP on life expectancy.

2.A multiple linear regression is suitable to analyze the data. The graph on the right shows the linearity in the data . Also, the colors in the plot represent the different groups (other, Africa, and OECD). We should definitely explore the interaction between log(PPGDP) and Group.

3. The following is a multiple linear regression model that will help us answer the questions we have regarding the relationship between the PPGDP and life expectancy.

y­­i =0 + ­1 log(PPGDP) + 2I(Group=Africa) + 3­­I(Group=OECD)+ β4­I(log(PPGDP)I(Group=Africa) + β5­I(log(PPGDP)I(Group=OECD) + i

where εi ~ N(0,σ2)

yi = the life expectancy for the ith country

β0 = the average life expectancy for a country that is not a member of either OECD or Africa and has a log(PPGDP) of 0.

β1 = the average increase in life expectancy as the log(PPGDP) increases by 1 for a country that is not a member of either OECD or Africa,

2 = the average change in life expectancy as you move from being a member of "other" to being a member of "Africa"

3 = the average change in life expectancy as you move from being a member of "Africa" to being a member of "OECD"

4 = the additional average increase in life expectancy as you increase log(PPGDP) by 1 for a country that is a member of Africa

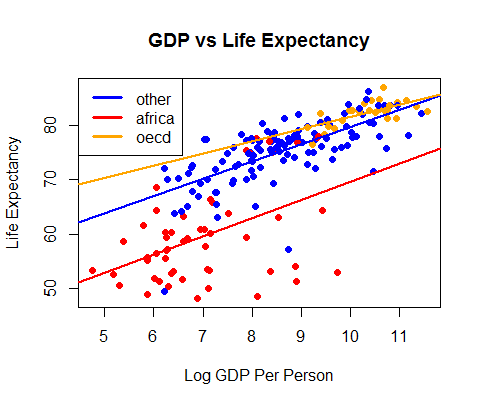
5 = the additional average increase in life expectancy as you increase log(PPGDP) by 1 for a country that is a member of OECD

This model assumes linearity, independence, normality, and equal variance.

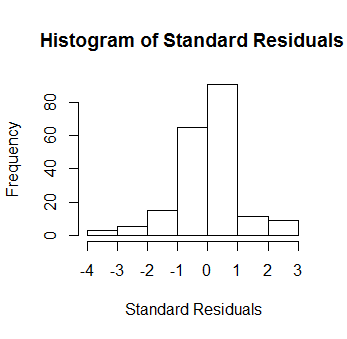
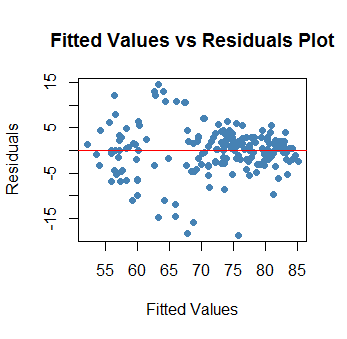
4. After fitting the model to the data, we get the following equation for life expectancy:

y­­i = 48.046 + ×log(PPGDP) + -11.8117×I(Group=Africa) + 11.1731×I(Group=OECD)+ .1655×I(log(PPGDP)I(Group=Africa) + -.9294×I(log(PPGDP)I(Group=OECD)

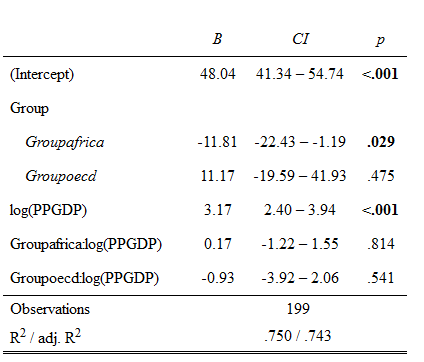
Also, we have plotted a regression line, with interactions on the data. It is shown below.



5. This model has an R2value of .7497, meaning that roughly 75% of the variation in Life Expectancy can be explained by this model. This is fairly high and indicates a good model fit. As stated earlier, this model assumes linearity, independence, normality, and equal variance. We have already commented on the linearity in the data, but we will also take a look at the other assumptions.

 Looking at the residuals versus predicted values below, the data appears to have equal variance across the line. It does not look perfect, but we are satisfied with this assumption. There are also no apparent patterns in this plot, so we will assume these data are independent. The Histogram of Residuals shows us that the errors are normally distributed with maybe a few outliers on the low end. Although these assumptions are not met perfectly, we will continue to use this model.

6. With a p-value of essentially 0, we can conclude that at least one of our predictor variables has a significant effect on life expectancy.

The table on the left shows all of the variables in the model with their corresponding coefficients and confidence intervals for those coefficients. For example, we are 95% confident that as log(PPGDP) increases by 1, the life expectancy will increase between 2.4 and 3.94.

On the basis of these intervals, this data does support the economists claim that countries with higher GDP have longer life expectancy.

7. We performed an F-test to determine if the interaction between log(PPGDP) and Group is significant. With a p-value of .7862, we cannot reject the null hypothesis that the model without the interaction is better. This means that the interaction is not actually significant. This means that the change in life expectancy as log(PPGDP) rises or falls will be the same for all countries, regardless of the group which they belong to.

## Appendix (Code)

setwd("~/3Fall2016/stat330/data")

life <-read.table("LifeExpectancy.txt", header=TRUE)

head(life)

### Choosing Base Line ####

unique(life$Group)

life$Group <- factor(life$Group, levels=c("other", "africa", "oecd"))

#### Exploring Data with various plots ####

plot(LifeExp~. , data=life)

### Plotting Transformed Data ####

my.colors <- c("blue", "red", "orange")

plot(log(life$PPGDP), life$LifeExp, main =

"GDP vs Life Expectancy", xlab= "Log GDP Per Person", ylab="Life Expectancy",

col=my.colors[life$Group], pch=19)

#### Exploring Boxplots (categorical data) ###

boxplot(life$LifeExp ~ life$Group, xlab="Group", ylab="Life Expectancy",

main="Life Expectancy by Group")

#### Building Multiple Linear Model ###

life.lm <- lm(LifeExp~ Group + log(PPGDP)+ Group:log(PPGDP), data=life)

#### Evaluating Assumptions ####

library(car)

avPlots(life.lm)

with(life.lm, plot(fitted.values, residuals, pch=19, col= "steelblue",

main="Fitted Values vs Residuals Plot", xlab="Fitted Values", ylab="Residuals"))

abline(h=0, col="red")

##### BP Test validates our homoskedicity ###

library(lmtest)

bptest(life.lm) ## P Value > 0.05 means data is homeskedastic ###

library(MASS)

stand.res <-stdres(life.lm)

hist(stand.res, main="Histogram of Standard Residuals", xlab=" Standard Residuals")

cooks <- cooks.distance(life.lm)

which(cooks > 4/length(stand.res))

ks.test(stand.res, "pnorm") ### Validates Normality ###

### R-squared and other coefficients ####

summary(life.lm)

sum.life <- summary(life.lm)

coef(sum.life)

sum.life$r.squared

### Running F-Test to test to see if interaction term is significant ###

reduced.lm <- lm(LifeExp ~ Group+log(PPGDP), data=life)

anova(life.lm, reduced.lm)

## Confidence Intervals ####

library(sjPlot)

sjt.lm(life.lm)

confint(life.lm, level=.95)

#### Plotting Regression Line to Fit Data ####

plot(log(life$PPGDP), life$LifeExp, main =

"GDP vs Life Expectancy", xlab= "Log GDP Per Person", ylab="Life Expectancy",

col=my.colors[life$Group], pch=19)

legend("topleft", legend=c("other", "africa", "oecd"),

col=c("blue", "red", "orange"), lty=1, lwd=3)

coef(life.lm)

### PPGDP and other ###

abline(coef(life.lm)[1], coef(life.lm)[4], col = "blue", lwd=2)

### PPGDP and africa ####

abline(sum(coef(life.lm)[c(1,2)]),sum(coef(life.lm)[c(4,5)]), col="red", lwd=2)

### PPGDP and oecd ###

abline(sum(coef(life.lm)[c(1,3)]),sum(coef(life.lm)[c(4,6)]), col="orange", lwd=2)