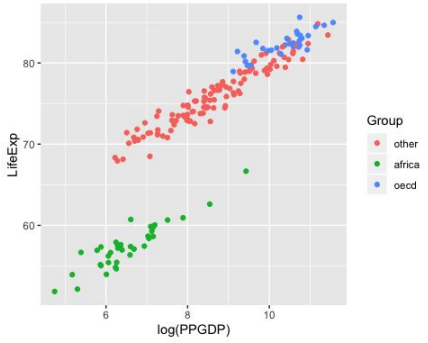
Elizabeth Mitchell
Stat 330

# Life Expectancy Analysis

We want to use this data to analyze if the economic well being of a country has a relationship with the life expectancies of the citizens of the country. We can create a statistical model of the relationship between these variables. The model will define a way to describe the relationship, for example we will know how much we expect life expectancies to increase with an increase in economic well being. Additionally we will be able to predict life expectancy given the economic well being of a country.




Based on the scatterplot of life expectancy compared to the log of PPGDP, we need to include an interaction in the model because each of the three groups follows a different line. This shows that the group that a country is in changes how the PPGDP affects life expectancy. Therefore we will include an interaction between log(PPGDP) and Group in our model.

The model we will use is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \varepsilon_i$

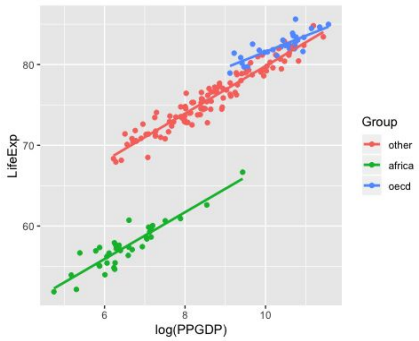
$\varepsilon \sim N(0, \sigma^2)$

$\beta_0$ is the life expectancy if log(PPGDP) is 0 and the group is "Other." $\beta_1$ is how much the life expectancy of a country will increase if all else is held constant and the log(PPGDP) goes up by one. $\beta_2$ is how much the the life expectancy of a country will increase if all else is held constant and the group is OECD. $\beta_3$ is how much the the life expectancy of a country will increase if all else is held constant and the group is Africa. $\beta_4$ is how much the life expectancy will go up if the log(PPGDP) increases by one and the group is OECD. $\beta_5$ is how much the life expectancy will go up if the log(PPGDP) increases by one and the group is Africa. $y_i$ is an individual estimate of the life expectancy. $x_{i1}$ is the estimate of an individual point of log(PPGDP). $x_{i2}$ is whether or not the group is OECD. $x_{i3}$ is whether or not the group is Africa. $\varepsilon_i$ is the error term and it is normally distributed.

By using this model we are assuming that there is a linear relationship, the data is independent, the residuals have a normal distribution and that there is equal variance.

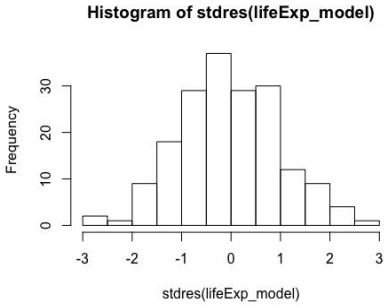
The fitted model for this data is

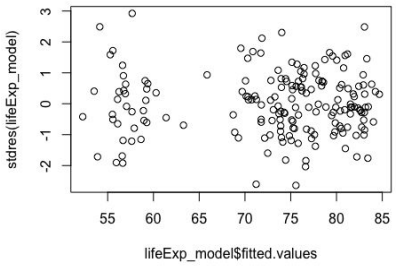
$$y_i = 50.424 + 2.939x_{i1} + -11.895x_{i2} + 11.292x_{i3} + -.041x_{i1}x_{i2} + -.953x_{i1}x_{i3} + \varepsilon_i$$




This model fits really well with an $R^2$ value of .987. This means that the model accounts for 98.7% of the variation in the distance from the points to the line that best fits them.

The assumption of linearity is met as we can see in the above scatterplot. The independence assumption is debatable, I would have to talk to an expert to know how one country is affected by another. For the time being we will assume independence. The assumption of normality is met as we can see in the histogram of the standardized residuals as shown below. Equal variance is also met as we can se from the plot of fitted values vs. residuals below.



The p-value for the overall F-test for the model is < 2.2e-16 so we can conclude that at least one of the factors included (PPGDP and Group) has a significant relationship with a country's life expectancy.

The 95% confidence interval for the intercept is (49.0171145, 51.8309469).

The 95% confidence interval for the slope of log(PPGDP) is (2.7772377, 3.1003980). This means that when all else is held constant we are 95% confident that as the log(PPGDP) increases by one on average a country's life expectancy will increase by between 2.78 and 3.1.

The 95% confidence interval for the slope of the group Africa is (-14.8068732, -8.9833514).

The 95% confidence interval for the slope of the group OECD is ( 4.9500705, 17.6339585).

The 95% confidence interval for the interaction of log(PPGDP) and group Africa is (-0.4606503, 0.3780884).

The 95% confidence interval for the interaction of log(PPGDP) and group OECD is (-1.5700033, -0.3353494).This means we are 95% confident that as the log(PPGDP) increases by one if the country is part of group OECD on average a country's life expectancy will decrease between 1.57 and .34.

The data support that as a country's GDP increases the life expectancy also increases since as the log of PPGDP increases since the confidence interval of the slope (2.7772377, 3.1003980) is positive, as log(PPGDP) increases so does life expectancy. However, since the interaction between log of PPGDP and the group OECD only contains negative values in the 95% confidence interval, if a country is in the OECD then as their PPGDP increases on average their life expectancy decreases (slightly.) To summarize, the data does support the economists claim but only conditionally.

The F test for the interaction between log(PPGDP) and Group has a p-value of .01083 indicating that the interaction is indeed significant. This confirms our interpretation that the way that an increase in PPGDP influences a country's average life expectancy is affected by which group the country belongs to.

Appendix:

```
library(ggplot2)
library(car)

## Loading required package: carData

library(MASS)

lifeExpect <- read.csv("LifeExp.txt", head = T, sep = " ")
head(lifeExpect)

##     Country Group   PPGDP LifeExp
## 1   Albania other  3677.2   75.31
## 2  Anguilla other 13750.1   79.19
```

```
## 3 Argentina other  9162.1   77.72
## 4   Armenia other  3030.7   74.54
## 5     Aruba other 22851.5   79.80
## 6 Australia  oecd 57118.9   83.37
```

```
lifeExpect$Group <- relevel(lifeExpect$Group, ref = "other")
```

```
ggplot(data=lifeExpect, mapping=aes(x=log(PPGDP), y=LifeExp, color = Group)) + geom_point()
```

```
lifeExp_model <- lm(LifeExp ~ log(PPGDP) + Group + log(PPGDP):Group, data = lifeExpect)
summary(lifeExp_model)
```

```
##
## Call:
## lm(formula = LifeExp ~ log(PPGDP) + Group + log(PPGDP):Group,
##     data = lifeExpect)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7772 -0.6729 -0.1000  0.6446  3.0438
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              50.42403    0.71286  70.734  < 2e-16 ***
## log(PPGDP)                2.93882    0.08187  35.896  < 2e-16 ***
## Groupafrica             -11.89511    1.47535  -8.063 1.13e-13 ***
## Groupoecd                11.29201    3.21337   3.514 0.000562 ***
## log(PPGDP):Groupafrica   -0.04128    0.21249  -0.194 0.846187
## log(PPGDP):Groupoecd     -0.95268    0.31279  -3.046 0.002680 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 175 degrees of freedom
## Multiple R-squared:  0.9865, Adjusted R-squared:  0.9861
## F-statistic:  2551 on 5 and 175 DF,  p-value: < 2.2e-16
```

```
ggplot(data=lifeExpect, mapping=aes(x=log(PPGDP), y=LifeExp, color = Group)) + geom_point() +
geom_smooth(method="lm", se=FALSE)
```

```
hist(stdres(lifeExp_model))
```

```
plot(lifeExp_model$fitted.values, stdres(lifeExp_model))
```

```
confint(lifeExp_model)
```

```
##                               2.5 %     97.5 %
## (Intercept)              49.0171145 51.8309469
## log(PPGDP)                2.7772377  3.1003980
## Groupafrica             -14.8068732 -8.9833514
## Groupoecd                 4.9500705 17.6339585
## log(PPGDP):Groupafrica   -0.4606503  0.3780884
## log(PPGDP):Groupoecd     -1.5700033 -0.3353494
```

```
reduced_model <- lm(LifeExp ~ log(PPGDP) + Group, data = lifeExpect)
anova(lifeExp_model, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: LifeExp ~ log(PPGDP) + Group + log(PPGDP):Group
## Model 2: LifeExp ~ log(PPGDP) + Group
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    175 195.12
## 2    177 205.48 -2   -10.357 4.6447 0.01083 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```