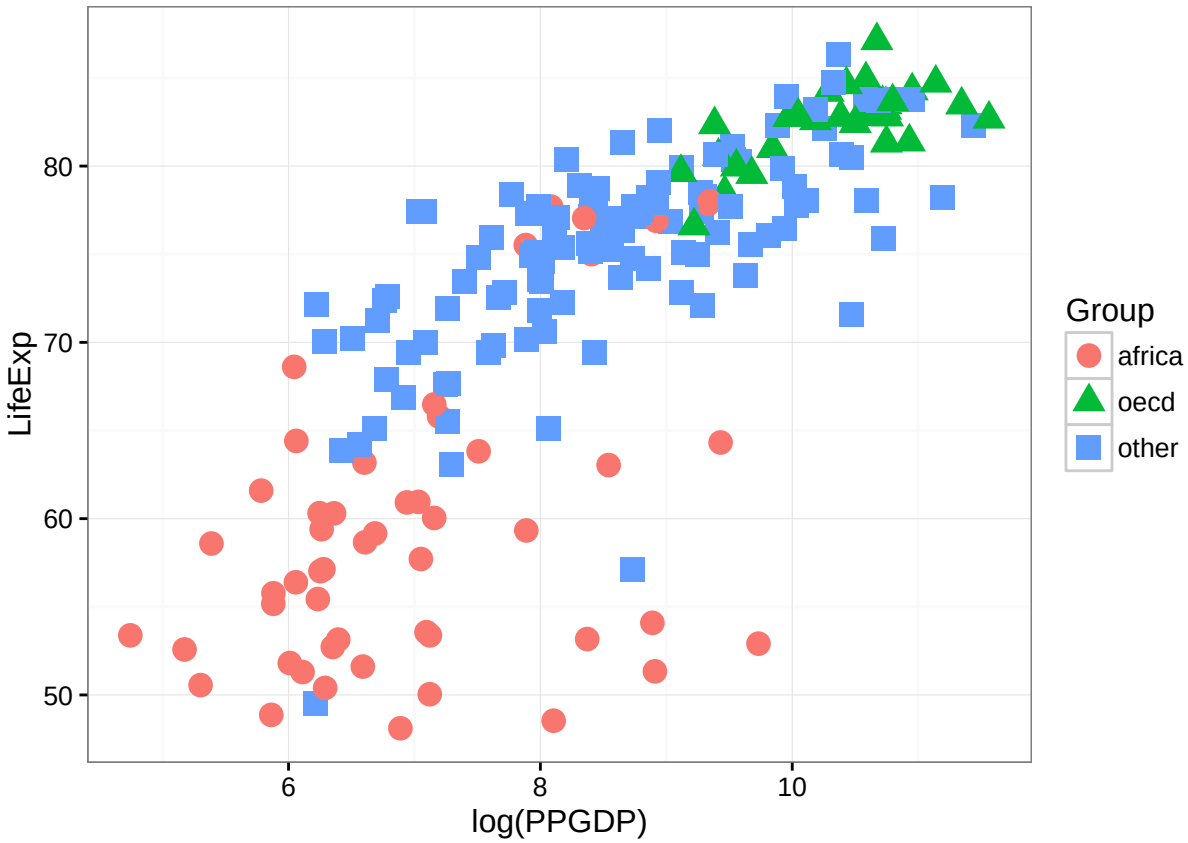
Life Expectancy

Cody Frisby

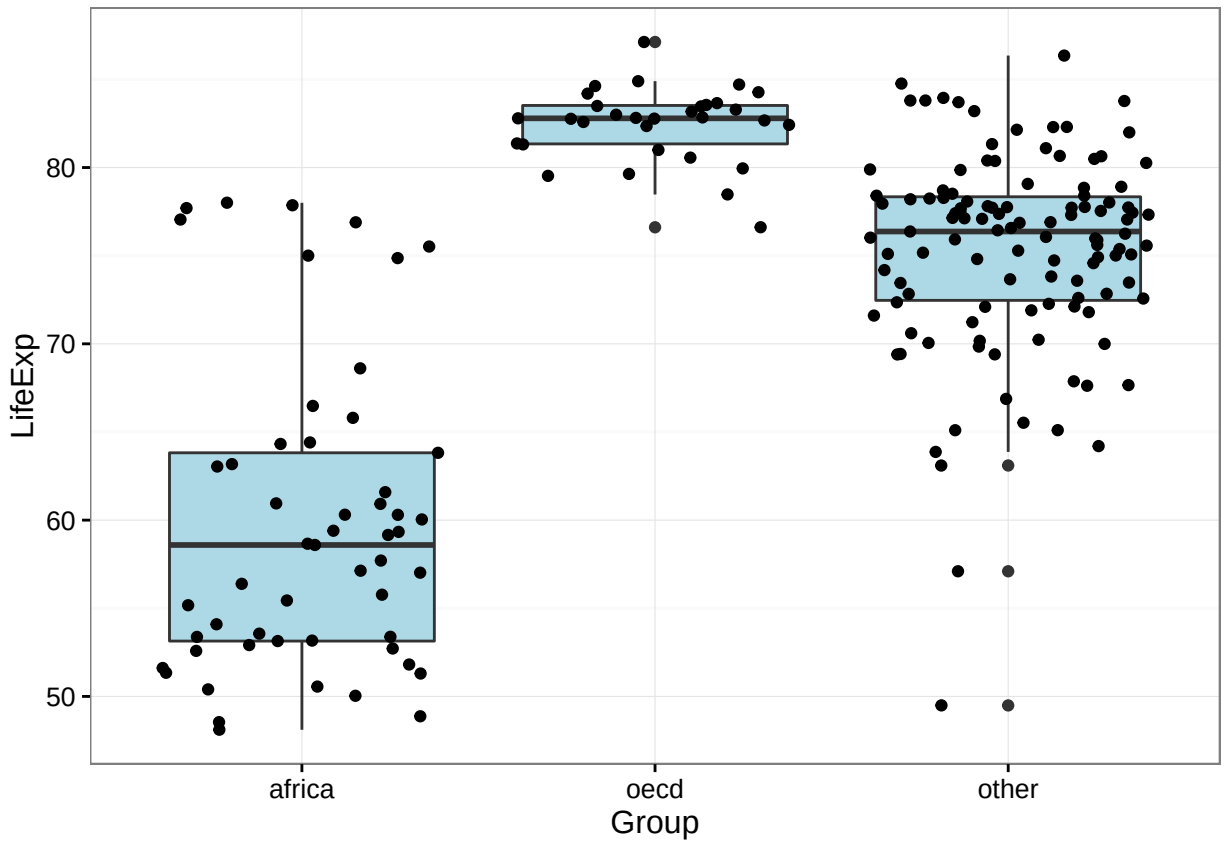
March 21, 2016

1) MLR and Variables to include

Here we plot the log of PPGDP by LifeExp with color/shape overlay of Group.



Now to look at box plots LifeExp by Group.



It appears that the indicator variables should be included. Group appears to have an influence on the life expectancy. There doesn't seem to be immediate need to include interaction terms in our model. Group and log(PPGDP) appear to be related but there doesn't appear to be an interaction between them. Also of note, group may not be needed at all, except **africa**, since most of the explanatory of **LifeExp** appears to be explained by **log(PPGDP)**.

2) Model

Note: $X_1 = log(PPGDP)$, $X_2 = 1$, if $Group =$ **africa**, $X_3 = 1$, if $Group =$ **oecd**. The model is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon$$

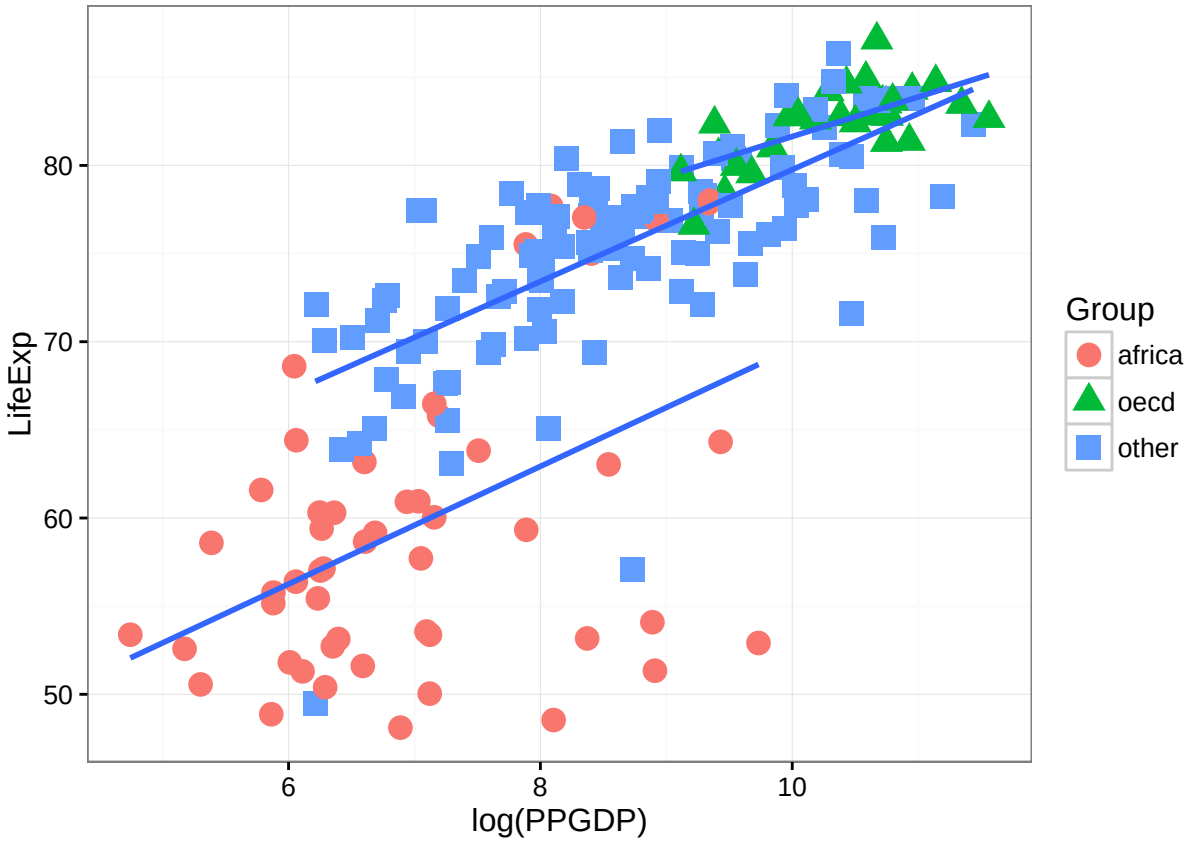
where β_0 is the expected life expectancy with no PPGDP and group equal to **other**, β_1 is the expected change in life expectancy given **log(PPDGP)** when country group is **other**, β_2 is the difference between an **africa** country and **other**, and β_3 is the difference between an **oecd** country and **other**, β_4 is the interaction between **log(PPDGP)** and *group* **africa**, β_5 is the interaction between **log(PPDGP)** and *group* **oecd**. These last two, β_4 & β_5 are simply additive, meaning how much additional or less *LifeExp* can we expect compared to *group* **other** when country is *Africa* or *oecd*, when considering **log(PPGDP)**. Here, we assume the errors from the above model are *normal* $(0, \sigma^2)$ and we assume a linear relationship between the predictors and the response **LifeExp**. This model will allow us to answer whether or not life expectancy is linked to PPGDG as well as quantify whether or not **OECD** countries have longer life expediencies.

3) Fit Model

Our fitted models' equation (from above) is

$$LifeExp = 48.0405584 + 3.1719725X_1 - 11.8117365X_2 + 11.1731029X_3 + 0.1655438X_1X_2 - 0.9294372X_1X_3$$

and here I display a scatter plot with our fitted models.



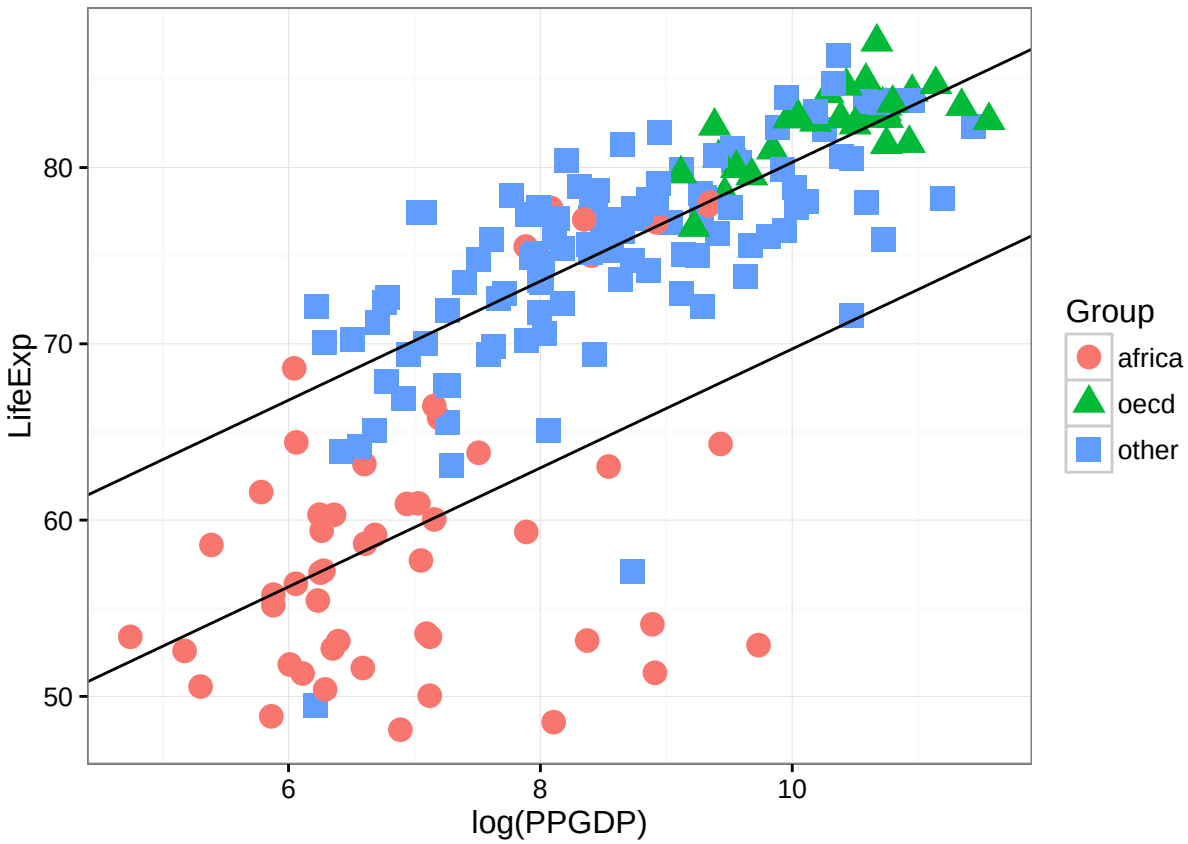
Where the top line is **oecd**, middle line is **other**, and the bottom is **africa**.

4) Reduced Model

The reduced model will be

$$LifeExp = 46.5657115 + 3.3728192X_1 - 10.5859227X_2$$

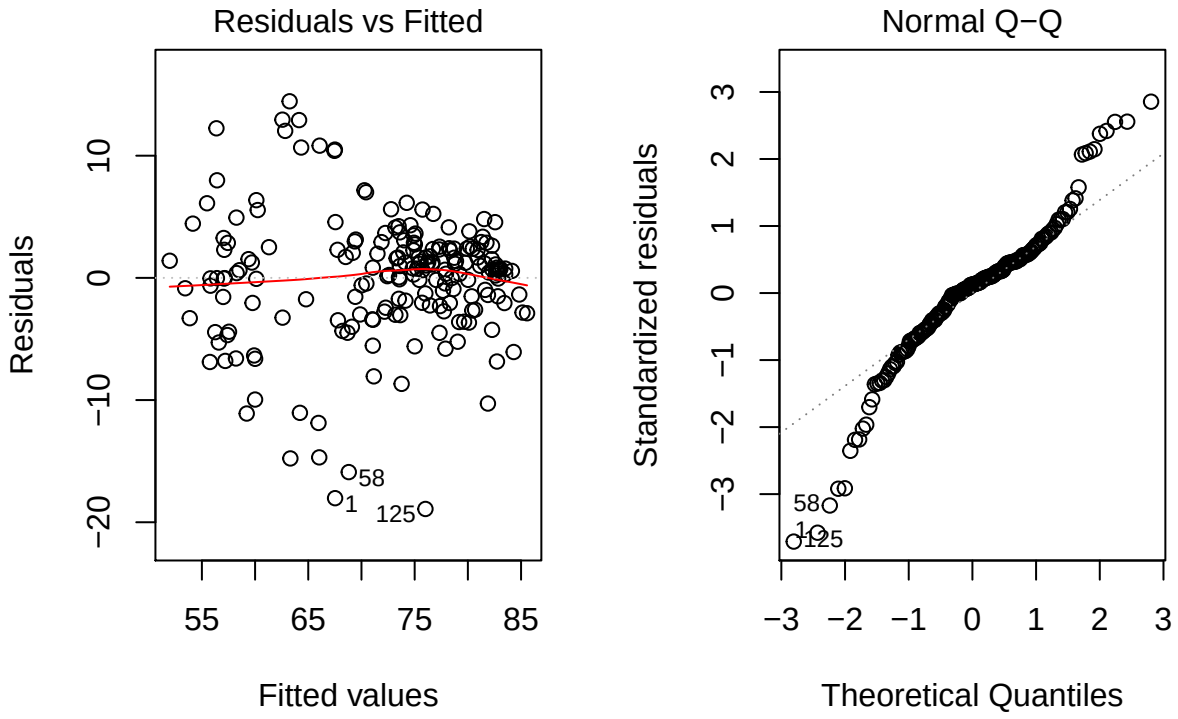
where, as before, X_1 is *log(PPGDP)* and $X_2 = 1$, if $Group = africa$. Here, we've removed all interactions and are now simply comparing groups *other* and *africa*. Displaying a plot below



where the top line is **other** and the bottom is **africa**.

5) Model Adequacy

Here, I take a look at some residual diagnostics for the above, reduced, model.



Of concern is the equality of variances assumption and the normality assumption. They aren't perfect, they don't need to be though, and we have transformed one of the predictor variables and this is often the case when we do a transformation. I think they are close enough and we will proceed with the above model. The reduced models R^2 is 0.7469595. This means that we can explain approximately 75% of the variation in life expectancy by knowing whether or not the country is from Africa and knowing the PPGDP of the country.

6) F test for the removal of covariates

$$F = \frac{\frac{SSE_{RM} - SSE_{FM}}{n-k}}{MSE_{RM}}$$

$$2.1873924 = \frac{\frac{5135.0054314 - 5077.6979223}{3}}{26.1990073}$$

The critical value for $F_{3,196:\alpha=0.05}$ is 2.6506765. So, we conclude that we have not adversely effected the model by removing the terms that we have removed.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
log(PPGDP)	1	12102.541	12102.54094	461.9465	0
africa	1	3055.673	3055.67292	116.6332	0
Residuals	196	5135.005	26.19901	NA	NA

Above, I display an ANOVA table of our reduced model. Included are F tests for the terms in the model. We can concluded that each of the included terms significantly effect *life expectancy*.

Below, I display 95% confidence intervals for each of our reduced model terms.
Intercept, which is interpretted as the expected *life expectancy* of countries in **Group** *oecd* or *other*, *on average*. *log(PPGDP)* is interpreted as the expected change in *life expectancy* per unit change in PPGDP, *on average*. And finally, **africa**, is interpreted as the expected life expectancy when comparted to *other* countries, *on average*.

	2.5 %	97.5 %
(Intercept)	41.559604	51.571819
log(PPGDP)	2.822888	3.922750
africa	-12.519029	-8.652817

7) Conclusions

We can answer YES to the question of whether or not a country's well being is linked with life expectancy. The inverval above, [2.8228882, 3.9227502], quantifies the effect that PPGDP has on life expectancy. For model assumptions to work we have taken the log of this value, but this interval does not include zero nor is it very close to zero. On average, we except that the log(PPGDP) has an effect on life expectancy by at least 2.8.

OECD countries will, on average, have higher life expectancies than **africa** countries but not nesecarrily **other** countries. If we consider **other** vs. **africa** where **other** is including **oecd** countries, then the interval displayed above for the line *africa* quantifies this difference for us. We can expect **other** countries to have a life expectancy of at least 8.6 points higher than *Africa* countries.

R code:

```
# read in the data from the excel file
life <- xlsx::read.xlsx
("~/Documents/MATH3710/ProblemSets/problem6/LifeExpectancy.xlsx", sheetIndex = 1)
# create indicator variables
life$africa <- 0
life$oecd <- 0
life[life$Group == "africa", "africa"] <- 1
life[life$Group == "oecd", "oecd"] <- 1
# plot with shape/color overlay
library(ggplot2)
g <- ggplot(data = life, aes(x=log(PPGDP), y=LifeExp))
g <- g + geom_point(aes(shape = Group, colour = Group), size = 4) + theme_bw()
g
b <- ggplot(life, aes(Group, LifeExp))
b <- b + geom_boxplot(fill = "lightblue") + geom_jitter() + theme_bw()
b
# below code creates a plot of predicted by observed as a way
# to visually see how "good" our model is.
fit.all <- lm(LifeExp ~ log(PPGDP)*africa + log(PPGDP)*oecd, data = life)
mod <- lm(life$LifeExp ~ fit.all$fitted.values)
rout <- list(paste('Fitted model: ', round(coef(mod)[1], 4), ' + ',
                   round(coef(mod)[2], 3), ' x', sep = ''),
             paste('R^2 == ', round(summary(mod)[['r.squared']], 3),
                   sep = ''))
#rout
df <- as.data.frame(cbind(x = fit.all$fitted.values, y = life$LifeExp))

fitplot <- ggplot(df, aes(x, y)) +
  geom_smooth(method = 'lm') + geom_point() + xlab("Predicted") +
  ylab("Observed") +
# need to change these x and y values below
```

```
  #geom_text(aes(x = 52, y = 80, label = rout[[1]]), hjust = 0) +
  geom_text(aes(x = 52, y = 78, label = rout[[2]]), hjust = 0,
            parse = TRUE)
fit2 <- lm(LifeExp ~ log(PPGDP)*Group, life) # not needed, same as fit.all
f <- formula(LifeExp~log(PPGDP)+africa+oecd+log(PPGDP)*africa+log(PPGDP)*oecd)
# less typing when we use this ^.
fit <- lm(LifeExp ~ log(PPGDP)+africa+oecd, data = life)
# no interactions ^.
b <- coef(fit.all) # model coefficients
# Slopes and intercepts from linear model
library(plyr)
coefs <- ddply(life, .(Group), function(df) {
  m <- lm(LifeExp ~ log(PPGDP), data = df)
  data.frame(a = coef(m)[1], b = coef(m)[2])
})
g <- ggplot(data = life, aes(x=log(PPGDP), y=LifeExp))
g <- g + geom_point(aes(shape = Group, colour = Group), size = 4) + theme_bw()
g <- g + geom_smooth(aes(group=Group), method = "lm", se=F, fullrange=F)
#g <- g + geom_abline(data=coefs, aes(intercept=a, slope=b))
g
# reduced model and coefs
fit.red <- lm(LifeExp ~ log(PPGDP) + africa, life)
b0 <- coef(fit.red)
# plot
df <- data.frame(b0 = c(b0[1], b0[1]+b0[3]), b1 = c(b0[2],b0[2]))
g <- ggplot(data = life, aes(x=log(PPGDP), y=LifeExp))
g <- g + geom_point(aes(shape = Group, colour = Group), size = 4) + theme_bw()
g <- g + geom_abline(data=df, aes(intercept=b0, slope=b1))
g
# plot model diagnostics
par(mfrow=c(1,2))
plot(fit.red, which = c(1,2))
# F test
sse.fm <- anova(fit.all)[[2]][6]
sse.rm <- anova(fit.red)[[2]][3]
mse.rm <- sse.rm/anova(fit.red)[[1]][3]
f.stat <- (sse.rm - sse.fm)/(mse.rm)
F0 <- qf(0.95, 3, 196)
f.p <- 1 - pf(f.stat, 3, 196)
```