

# LifeExpectancy

Jillian Maw

3/3/2022

## Homework Analysis #5 - Life Expectancy

Life expectancy is often used as an indicator for a well-being of a country. Macroeconomists often speculate that the life expectancy of linked with the economic well-being of a country. To test this hypothesis, the `LifeExpectancy.txt` dataset contains the following information:

Variable Name	Description
Country	Country name
Group	Is the country a member of OECD, Africa, or other?
PPGDP	Per person GDP

The **Group** variable indicates if the country is a member of the Organisation for Economic Co-operation and Development, an international think tank charged with promoting policies that will improve global social and economic well-being (**FOR THIS ASSIGNMENT: use “other” as the baseline category**). Macroeconomists also hypothesize that OECD members will have longer life expectancy.

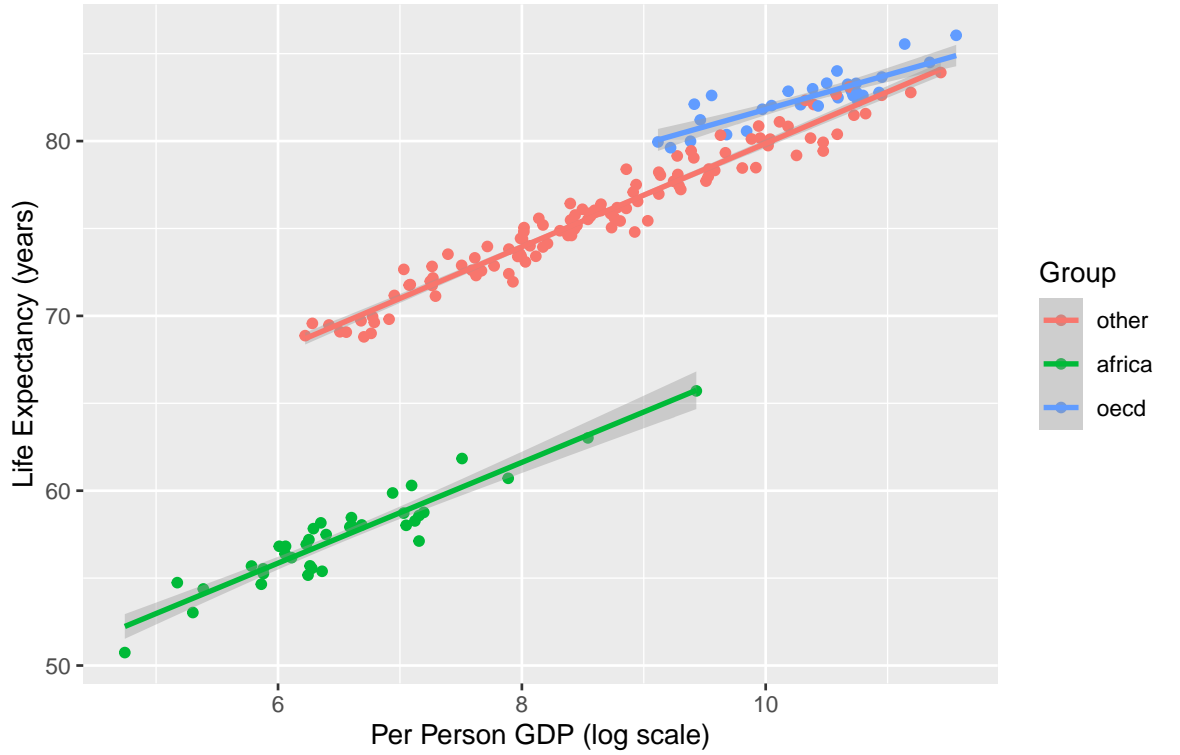
For each of the following questions, assume that your audience are political strategists working for the OECD. Please attach your clearly commented code (R or Python) as an appendix.

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

The economic well-being of a country could be determined with an indicator such as life expectancy. Macroeconomists wish to test this hypothesis and the hypothesis that if a country is a member of the Organisation for Economic Co-operation and Development, the OECD, it will have a longer life expectancy. Statistical modeling can prove if there is a linear relationship (and what is it) between per person global domestic product (PPGDP), an economic measurement, and life expectancy in years. If this relationship does, in fact, exist at a significant level, statistical modeling can also provide a linear relationship model that can provide an estimate of a country’s life expectancy, given their PPCDP. It could be determined then, if that country is a member of the OECD, if they also tend to have longer life expectancies.

2. Set the baseline of the **group** variable to **other**. Provide a plot of the life expectancy vs.  $\log(\text{PPGDP})$  where you use different plotting symbols (or colors) for countries of each group type. From this plot, provide a discussion on whether or not interactions between  $\log(\text{PPGDP})$  and **Group** should be included in the

Per Person GDP vs Life Expectancy: Transformed Data



model.

Based on the graph above, we see that there are different slopes for each group type's line of regression. It would be appropriate, therefore, to include the interaction between the log of Per Person Gross Domestic Product and Life Expectancy to account for the differences in these group type's linear relationship demonstrating that as the log of Per Person Gross Domestic Product increases in a group type's member countries, their life expectancy also increases.

3. Write out (in mathematical form with Greek letters) a MLR model that would help answer the questions you stated in #1 *and* includes an interaction term between  $\log(\text{PPGDP})$  and **Group**. Clearly state any assumptions you are using in your model. Provide an interpretation of each mathematical term (variable or parameter) included in your model (**be careful to interpret the interaction terms correctly**). Using the mathematical form, discuss how your model, after fitting it to the data, will be able to answer the questions in this problem.

A multiple linear regression model that would help answer the questions stated above, using Greek letters, mathematical notation, and the interaction between the log of per person Gross Domestic Product and group, could be the following:

$$y_i = \beta_0 + \beta_1 I(\text{Group} = \text{Africa}_i) + \beta_2 I(\text{Group} = \text{OECD}_i) + \beta_3 (\log(\text{PPGDP}_i)) +$$

$$\beta_4 I(\text{Group} = \text{Africa}_i)(\log(\text{PPGDP}_i)) + \beta_5 I(\text{Group} = \text{OECD}_i)(\log(\text{PPGDP}_i)), \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

Above,  $y$  represents the expected life expectancy of a country  $i$ . The terms following  $I$ , such as  $(\text{Group} = \text{Africa}_i)$  and  $(\text{Group} = \text{OECD}_i)$ , which represent whether a country is or is not a member of the OECD, Africa, or other group, and  $(\log(\text{PPGDP}_i))$ , which represents the  $i$ th country's log value of the per person Gross Domestic Product, are the explanatory variables of this model. The group variable is categorical, and would be substituted with 0 for not being part of that group and 1 for yes being part of that group. The per person Gross Domestic Product variable is a quantitative variable.  $\beta_0$  represents the intercept, or when all the other measurements are zero and the group membership is "other", then  $y$  is  $\beta_0$ , on average; this is a next-to-useless interpretation, however, and it is better to think of  $\beta_0$  as the base, average life

expectancy in years of any human being in the group “other”, our baseline for membership. Next, we clarify the representations of the coefficients of this model.  $\beta_1$  represents whether or not a country is part of the group of Africa or not,  $\beta_2$  represents whether or not a country is part of the OECD,  $\beta_3$  represents the log of a country’s Per Person Gross Domestic Product,  $\beta_4$  represents the interaction between the log of a country’s Per Person Gross Domestic Product and a country being in the group part of Africa,  $\beta_5$  represents the interaction between the log of a country’s Per Person Gross Domestic Product and a country being in the group OECD, and  $\sigma^2$  represents the unbiased variance of the model, or the average distance of life expectancy in years from the multiple linear regression model’s “line” of estimation.

Since “interaction” is not the most intuitive way to think of this model, we can break this down by “line of regression” in the model. There are three groups in the model: other, Africa, and OECD. They have the following slopes:

$$y_i = \begin{cases} \beta_0 + \beta_1(\log(\text{PPGDP}_i)) & \text{if Group}_i = \text{“other”} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4)(\log(\text{PPGDP}_i)) & \text{if Group}_i = \text{“Africa”} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5)(\log(\text{PPGDP}_i)) & \text{if Group}_i = \text{“OECD”} \end{cases}$$

What these regression lines mean depends on the group the country  $i$  is part of. If the country  $i$  is part of the “other” group, then for a  $\log(\text{PPGDP}_i)$  of 0, I expect the life expectancy of that country to be  $\beta_0$  years. Additionally, as  $\log(\text{PPGDP}_i)$  increases by 1, we expect the life expectancy for countries in the “other” group to increase by  $\beta_1$  years. If the country  $i$  is part of the “Africa” group, then for a  $\log(\text{PPGDP}_i)$  of 0, we expect the life expectancy of that country to be  $\beta_0 + \beta_2$  years. Also, as  $\log(\text{PPGDP}_i)$  increases by 1, we expect the life expectancy for countries in Africa to increase by  $\beta_1 + \beta_4$  years. If the country  $i$  is part of the “OECD” group then for a  $\log(\text{PPGDP}_i)$  of 0, we expect the life expectancy of that country to be  $\beta_0 + \beta_3$  years. Finally, as  $\log(\text{PPGDP}_i)$  goes up by 1 for OECD countries, we expect the life expectancy increases by  $\beta_1 + \beta_5$  years.

4. Fit your model in #3 to the Life Expectancy data and summarize the results by displaying the fitted model in equation form (do NOT just provide a screen shot of the R or Python output). Plot your fitted regression line (with interactions) on the scatterplot of life expectancy vs.log(PPGDP).

To fit our model above with the estimated values to the Life Expectancy data, we clarify that these are estimated values for the multiple linear regression model, or as we see below, estimated coefficients of best fit:

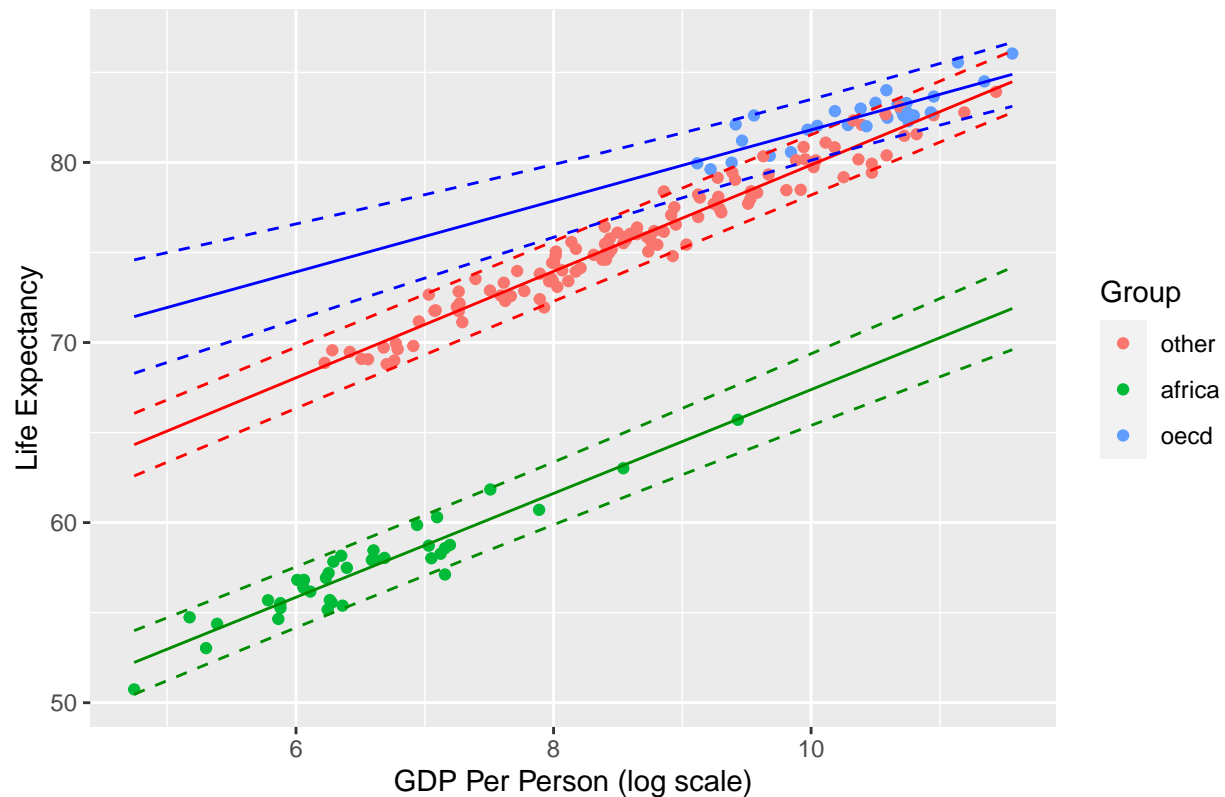
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 I(\text{Group} = \text{Africa}_i) + \hat{\beta}_2 I(\text{Group} = \text{OECD}_i) + \hat{\beta}_3 (\log(\text{PPGDP}_i)) + \hat{\beta}_4 I(\text{Group} = \text{Africa}_i)(\log(\text{PPGDP}_i)) + \hat{\beta}_5 I(\text{Group} = \text{OECD}_i)(\log(\text{PPGDP}_i)), \epsilon_i \stackrel{iid}{\sim} N(0, \hat{\sigma}^2)$$

Which we then rewrite as:

$$\hat{y}_i = 50.3118 + -11.7484 I(\text{Group} = \text{Africa}_i) + 11.7717 I(\text{Group} = \text{OECD}_i) + 2.9552 I(\log(\text{PPGDP}_i)) + -0.073 I(\text{Group} = \text{Africa}_i)I(\log(\text{PPGDP}_i)) + -0.9826 I(\text{Group} = \text{OECD}_i)I(\log(\text{PPGDP}_i)), \epsilon_i \stackrel{iid}{\sim} N(0, 0.7052).$$

We can see the estimated values of the multiple linear regression in the scatterplot below, where the lines of regression are split by group: red for “other”, green for “africa”, and blue for “oecd”. The dashed lines are the prediction range of the lines of regression.

## GDP Per Person vs Life Expectancy (With Regression)

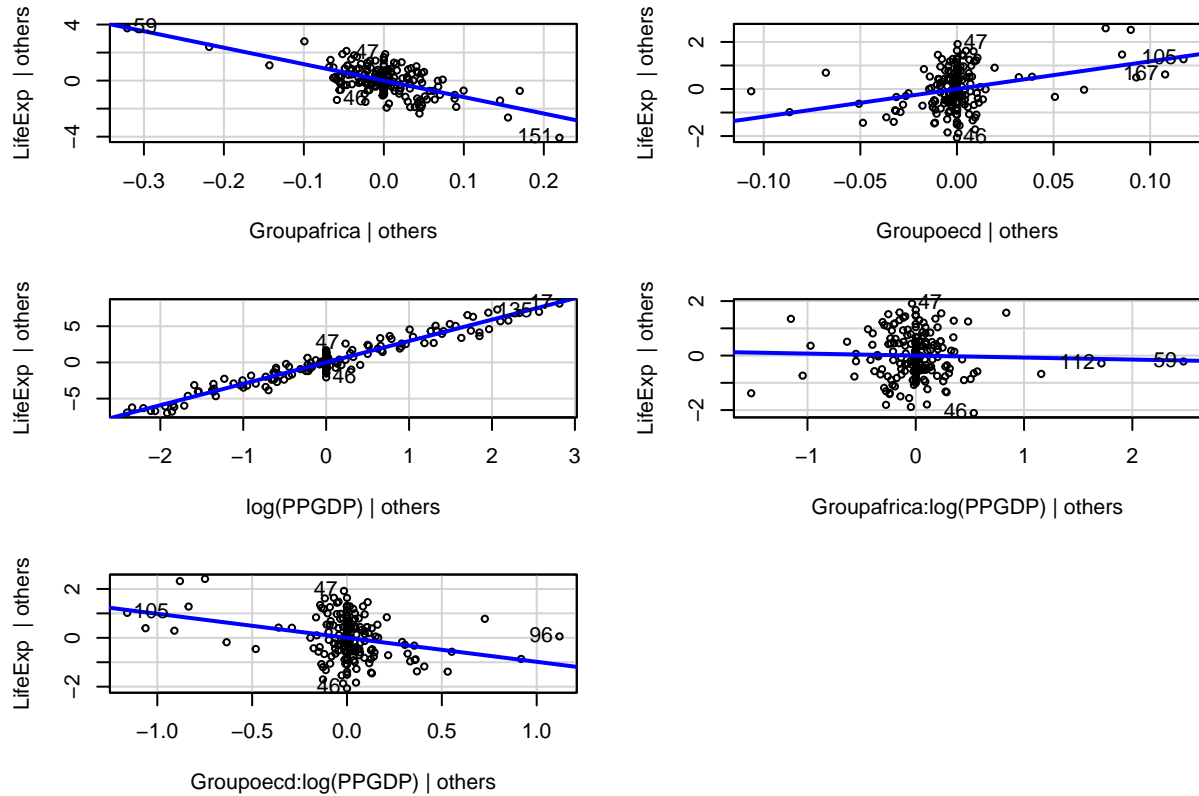


5. Assess the fit of your model and justify your model assumptions using appropriate graphics or summary statistics. Provide discussion of your assessment of the model fit and assumptions on the level of your target audience (e.g. interpret your model  $R^2$ ).

For this multiple linear regression model, we assumed that the regressions are linear, independent, all follow a Normal distribution, and that the data has equal variance about the regression lines. We will prove this with the graphics and summary statistics below.

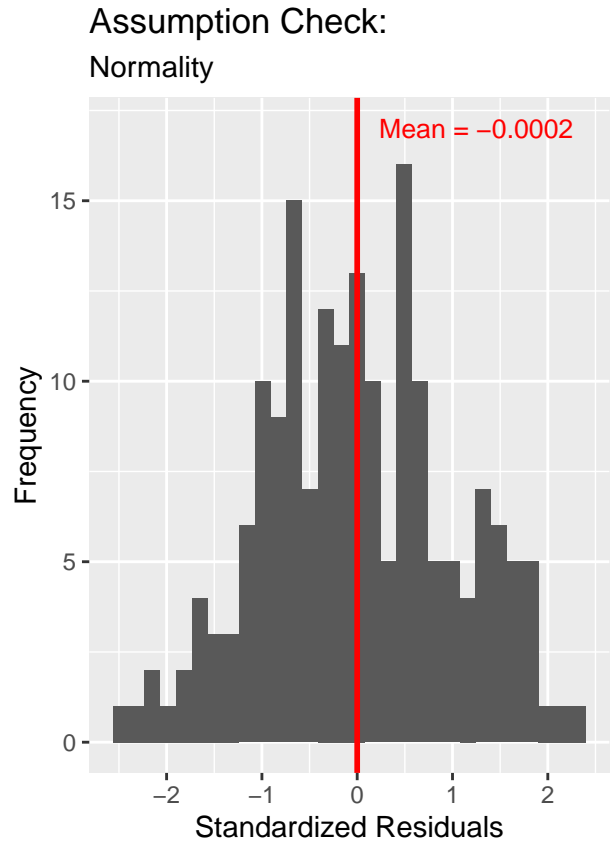
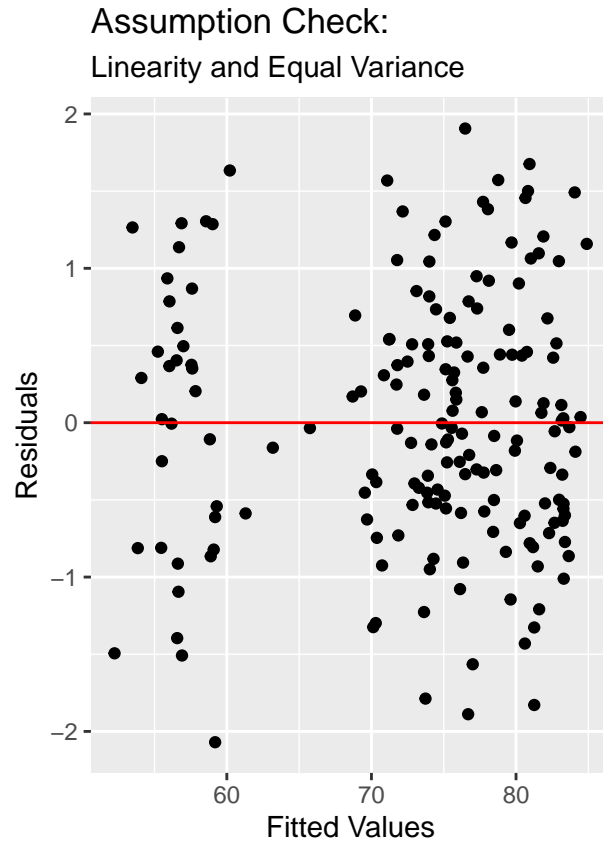
First, we will look at the Added-Variable Plots below. The graphs all seem to show clustering data at some central point, with outliers possibly affecting the slope, in that they pull the slope through a cluster of points in one direction over another. However, there are no patterns in the plots showing that other graphical line would be appropriate (e.g. exponential), so we can assume the linearity assumption is met. We will double check the linearity with the soon-to-be shown fitted values versus standardized residuals plot after the Added-Variable Plots.

## Added-Variable Plots



We can assume the data is independent because one country's life expectancy should not affect another country's life expectancy. One country having a high life expectancy should not have an effect on another country also having a high or low life expectancy. Citizens in one country are not changing the life expectancy of another country to such an extent it should affect our data.

Finally, for our last graphics, we check the normality and equal variance assumptions. We will prove the equal variance assumption (and reaffirm the linearity assumption) by checking a plot comparing fitted values (predicted value for life expectancy in the dataset) to the residuals (the difference between the observed and a predicted life expectancies) in a scatterplot below. Since we can see a constant variance with a lack of patterns, the data has equal variance and that assumption seems to be proven correct. Next, we will prove the data follows a Normal distribution by plotting the standardized residuals (residuals transformed to show their difference from the data's mean, if the mean was 0) on a histogram. Below, we see the histogram shows a generally Normal distribution, so we assume the normality assumption is also met.



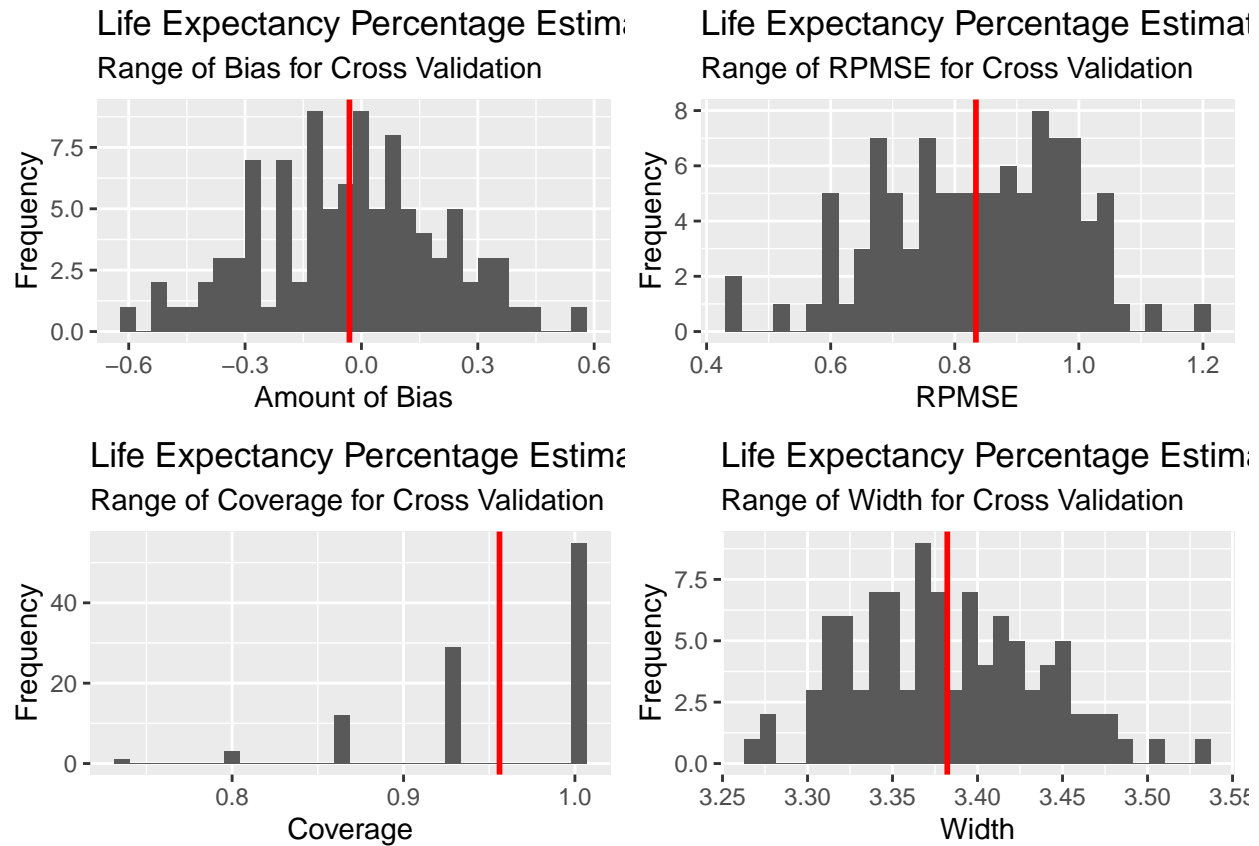
To further prove we meet the Normality assumption for the model, we conducted a One-sample Kolmogorov-Smirnov test, also called a KS-test, and a Jarque-Bera test for normality, also called a JB-test. These tests conduct hypothesis tests on whether or not a data set follows a Normal distribution or not. For the KS-test, the null hypothesis is that the data comes from a Normal distribution, while the alternative hypothesis is that the data does *not* come from a Normal distribution. For the JB-test, the null hypothesis is that the data's distribution is not skewed, whereas the alternative hypothesis is that the data's distribution *is* skewed. We set the p-value to be 0.05 for both tests, to prove significance. The KS-test produced a p-value of 0.9222, so we failed to reject the null hypothesis. The JB-test for normality produced a p-value of 0.363, so we failed to reject the null hypothesis for both tests. We accept that we have the normality assumption met for our data.

To further prove the Equal Variance assumption for our model, we conducted a Breusch-Pagan test, or BP-test. Checking the scatterplot of fitted values versus residuals above, it appears that we have a model with linearity and mostly equal variance, so we can proceed with the BP-test. The BP-test conducts hypothesis tests on whether or not a data set has homoskedasticity, or equal variance. The null hypothesis is that the data has homoskedasticity, while the alternative hypothesis is that the data has *heteroskedasticity*. We set the p-value to be 0.05, to prove significance. The test produced a p-value of 0.4192, so we failed to reject the null hypothesis. We accept that we have the Equal Variance assumption met for our data.

We conducted a cross validation procedure 100 times on 15 randomly selected observations of the data, newly selected each time, and calculated an average bias of -0.031. This means our predictions are, on average, slightly lower than the true average of life expectancy. We also calculated an average Root Predictive Mean Square Error of 0.8341, which means our predictions are off, on average, 0.8341 years. Considering the range of years of life expectancy is between 50.74 and 86.05 years, this seems a reasonable amount of error. To see how far our predictions ranged, we calculated the width to be 3.3822 years of life expectancy, on average. In addition, the coverage, or the percentage of prediction intervals that contain the true average of years of life expectancy, to be 0.956. Below the following paragraph are graphs showing in greater detail the results of

the cross validation procedures, with the red lines representing the mean values relative to the results shown.

In addition to the above cross validation results (and below graphical representation of the results), the percent of variability in years of life expectancy explained by the covariate variables in the adjusted multiple linear regression model above in #4 is  $R^2$ , 99.15. With all of these results and summary statistics, we feel it safe to say that our predictions are rather good, relative to the original spread of the response variable life expectancy in years.



6. Perform an  $F$ -test to determine if any of your predictors (covariates) significantly affect life expectancy. Construct 95% confidence intervals for each term (including the interactions). Interpret the confidence intervals for  $\log(\text{PPGDP})$  and  $\log(\text{PPGDP}):\text{Group}["\text{OECD}"]$ . Do the data support the economists claims that countries with higher GDP have longer life expectancy?

```
##
## Call:
## lm(formula = LifeExp ~ Group + log(PPGDP) + log(PPGDP):Group,
##     data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06994 -0.58491 -0.03349  0.51896  1.90535
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    50.31182    0.56693   88.744 < 0.0000000000000002 ***
## Groupafrica    -11.74840    1.17333  -10.013 < 0.0000000000000002 ***
```

```
## Groupoecd          11.77172    2.55557    4.606          0.00000786 ***
## log(PPGDP)         2.95517    0.06511   45.387 < 0.0000000000000002 ***
## Groupafrica:log(PPGDP) -0.07302    0.16899   -0.432          0.666211
## Groupoecd:log(PPGDP)  -0.98258    0.24876   -3.950          0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8398 on 175 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9913
## F-statistic:  4100 on 5 and 175 DF,  p-value: < 0.00000000000000022

##                2.5 %    97.5 %
## (Intercept)      49.1929106 51.4307297
## Groupafrica     -14.0641022 -9.4327004
## Groupoecd         6.7280204 16.8154184
## log(PPGDP)        2.8266641  3.0836710
## Groupafrica:log(PPGDP) -0.4065395  0.2605029
## Groupoecd:log(PPGDP)  -1.4735332 -0.4916225
```

By the  $F$ -test results above, we see that our  $F$ -test gives a resulting p-value of 0.00000000000000022, or so much below a significance level of 0.05, as to cause us to fail to reject the null hypothesis and determine that our predictors (covariates), of group, log of PPDP, and the interaction between them, significantly affect life expectancy of a country.

From our confidence intervals created above, we state we are 95% confident that the true effect of  $\log(PPGDP)$  on life expectancy will be, on average, between 2.8267 and 3.0837 years. We are 95% confident that the effect of the interaction between  $\log(PPGDP)$  and a group of “OECD” will be, on average, between -1.4735 and -0.4916 years. The first of these intervals supports the economists’ claims that countries with higher GDP have longer life expectancy because it shows that as GDP per person increases, the life expectancy increases. As for the second confidence interval, we cannot claim this because of the decreasing life expectancy from the interaction. Our best guess is that the interaction is tempered to accuracy by the decrease in life expectancy.

7. Perform an  $F$ -test to determine if the interaction between  $\log(PPGDP)$  and **Group** is significant. Explain how this significance (or lack thereof) affects the interpretation of the effect of  $PPGDP$  on life expectancy.

After performing an  $F$ -test, to determine if the interaction between  $\log(PPGDP)$  and **Group** is significant at the p-level of 0.05, we determine that the resulting p-value is well below this level at 0.0005682. This proves that the interaction is significant and should be included in our multiple linear regression model for accuracy of life expectancy predictions.

## Appendix of Code

```
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(ggplot2) #for professional looking graphics
library(GGally) #for ggpairs
library(MASS) #for standardized residuals
library(normtest) #for JB-test
library(lmtest) #for BP-test
# library(SciViews) #maybe for correlation and covariance matrices?
library(car) #for variance inflation factors & added-variable plots
```



```

library(bestglm) #for variable selection procedures
library(knitr) #for pretty tables with kable
# library(kableExtra) #for if kable needs to be landscape
library(gridExtra) #for making grids on same row
options(scipen = 999) #for preventing scientific notation
life = read.table("~/R programming/STAT_330/LifeExpectancy3.txt", sep = ' ', header = TRUE)
life$Group <- as.factor(life$Group)
life$Country <- as.character(life$Country)
# levels(life$Group)
# nlevels(life$Group)
# table(life$Group)
life$Group <- relevel(life$Group, ref = "other")
lifeplot <- ggplot(data=life, mapping=aes(y=LifeExp, x=log(PPGDP), color = Group)) +
  geom_point() +
  xlab('Per Person GDP (log scale)') +
  ylab('Life Expectancy (years)') +
  ggtitle('Per Person GDP vs Life Expectancy: Transformed Data') +
  geom_smooth(method="lm")
suppressMessages(print(lifeplot))
#yes b/c diff slopes
log.ppgdp.group <- interaction(life$Group, log(life$PPGDP))
#do I need to adjust for log(PPGDP)?
#What is a Morgan-Tatar search?
#"Morgan-Tatar search since factors present with more than 2 levels."
vs.res <- bestglm(life[, -1], IC="AIC", method="exhaustive", TopModels=15)
#Maybe take out Country?
life.mlr <- lm(LifeExp~Group+log(PPGDP)+log(PPGDP):Group, data = life)
#create table of coefficient estimates with kable
life.beta.0 <- round(as.numeric(coef(life.mlr)["(Intercept)"]), digits = 4)
life.beta.1 <- round(as.numeric(coef(life.mlr)["Groupafrica"]), digits = 4)
life.beta.2 <- round(as.numeric(coef(life.mlr)["Groupoecd"]), digits = 4)
life.beta.3 <- round(as.numeric(coef(life.mlr)["log(PPGDP)"]), digits = 4)
life.beta.4 <- round(as.numeric(coef(life.mlr)["Groupafrica:log(PPGDP)"]), digits = 4)
life.beta.5 <- round(as.numeric(coef(life.mlr)["Groupoecd:log(PPGDP)"]), digits = 4)
life.var <- round(sigma(life.mlr)^2, digits = 4)
log.seq <- seq(min(life$PPGDP), max(life$PPGDP), length=1000) #max(log(life$PPGDP))=11.56262
#data.frame(Experience=experience.seq, Education="HS", Manager="No")
log.df <- data.frame(PPGDP=log.seq, Group=rep("other", 1000))
log2.df <- data.frame(PPGDP=log.seq, Group=rep("africa", 1000))
log3.df <- data.frame(PPGDP=log.seq, Group=rep("oecd", 1000))
log.pred <- predict.lm(life.mlr, newdata=log.df, interval="prediction")
log2.pred <- predict.lm(life.mlr, newdata=log2.df, interval="prediction")
log3.pred <- predict.lm(life.mlr, newdata=log3.df, interval="prediction")
log.seq <- log(log.seq)
#scatterplot with lines
bigplot <- ggplot() +
  geom_point(data=life, mapping=aes(y=LifeExp, x=log(PPGDP), color = Group)) +
  geom_line(mapping=aes(x=log.seq, y=log.pred[, 'lwr']), color="red", linetype = 2) +
  geom_line(mapping=aes(x=log.seq, y=log.pred[, 'fit']), color="red") +
  geom_line(mapping=aes(x=log.seq, y=log.pred[, 'upr']), color="red", linetype = 2) +
  geom_line(mapping=aes(x=log.seq, y=log2.pred[, 'lwr']), color="green4", linetype = 2) +
  geom_line(mapping=aes(x=log.seq, y=log2.pred[, 'fit']), color="green4") +
  geom_line(mapping=aes(x=log.seq, y=log2.pred[, 'upr']), color="green4", linetype = 2) +

```

```

geom_line(mapping=aes(x=log.seq, y=log3.pred[, 'lwr']), color="blue", linetype = 2) +
geom_line(mapping=aes(x=log.seq, y=log3.pred[, 'fit']), color="blue") +
geom_line(mapping=aes(x=log.seq, y=log3.pred[, 'upr']), color="blue", linetype = 2) +
xlab('GDP Per Person (log scale)') +
ylab('Life Expectancy') +
ggtitle('GDP Per Person vs Life Expectancy (With Regression)')
suppressMessages(print(bigplot))
#linearity
avPlots(life.mlr, ask=FALSE)
#equal variance
fit.vs.resids <- ggplot(life, aes(x=life.mlr$fitted.values, y=life.mlr$residuals)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Assumption Check:',
    subtitle = 'Linearity and Equal Variance') +
  geom_hline(yintercept = 0, col = "red", lwd = 0.5)

#Normality
std.resids <- stdres(life.mlr)
hline <- round(mean(std.resids), digits = 4)
life.freq <- ggplot() +
  geom_histogram(mapping=aes(x=std.resids)) +
  xlab('Standardized Residuals') +
  ylab('Frequency') +
  ggtitle('Assumption Check:', subtitle = 'Normality') +
  geom_vline(xintercept = mean(std.resids), col = "red", lwd = 1) +
  annotate("text", x = hline + 1.25, y = 17,
    label = paste("Mean =", hline), col = "red", size = 3.5)
suppressMessages(grid.arrange(fit.vs.resids, life.freq, nrow=1))
ks.test(std.resids, "pnorm")
jb.norm.test(std.resids)
bptest(life.mlr)
#cross validation
set.seed(87) #set seed for reproducibility
n.cv <- 100 #Number of CV studies we'll run
bias <- rep(NA, n.cv) #n.cv empty biases (one for each CV)
RPMSE <- rep(NA, n.cv) #n.cv empty RPMSE (one for each CV)
coverage <- rep(NA, n.cv) #n.cv empty coverage (one for each CV)
width <- rep(NA, n.cv) #n.cv empty width (one for each CV)
n.test <- 15 #How big my test set is
for(i in 1:n.cv){
  test.obs <- sample(1:nrow(life), n.test)
  test.set <- life[test.obs,]
  train.set <- life[-test.obs,]

  train.lm <- lm(LifeExp~Group+I(log(PPGDP))+I(log(PPGDP)):Group, data=train.set)
  test.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

  bias[i] <- mean(test.preds[,1] - test.set$LifeExp)
  RPMSE[i] <- sqrt(mean((test.preds[,1] - test.set$LifeExp)^2))
  coverage[i] <- mean((test.preds[,2] < test.set$LifeExp) &
    (test.preds[,3] > test.set$LifeExp))
}

```

```

width[i] <- mean(test.preds[,3] - test.preds[,2])
}

life.stddev <- round(sigma(life.mlr), digits = 4)
mean.bias <- round(mean(bias), digits = 4)
mean.RPMSE <- round(mean(RPMSE), digits = 4)
mean.coverage <- round(mean(coverage), digits = 4)
mean.width <- round(mean(width), digits = 4)
life.r2 <- round(summary(life.mlr)$r.squared, digits = 4)
CV.bias <- ggplot() +
  geom_histogram(mapping=aes(x=bias)) +
  xlab('Amount of Bias') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of Bias for Cross Validation') +
  geom_vline(xintercept = mean.bias, col = "red", lwd = 1)

CV.RPMSE <- ggplot() +
  geom_histogram(mapping=aes(x=RPMSE)) +
  xlab('RPMSE') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of RPMSE for Cross Validation') +
  geom_vline(xintercept = mean.RPMSE, col = "red", lwd = 1)

CV.coverage <- ggplot() +
  geom_histogram(mapping=aes(x=coverage)) +
  xlab('Coverage') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of Coverage for Cross Validation') +
  geom_vline(xintercept = mean.coverage, col = "red", lwd = 1)

CV.width <- ggplot() +
  geom_histogram(mapping=aes(x=width)) +
  xlab('Width') +
  ylab('Frequency') +
  ggtitle('Life Expectancy Percentage Estimation:',
          subtitle = 'Range of Width for Cross Validation') +
  geom_vline(xintercept = mean.width, col = "red", lwd = 1)

suppressMessages(grid.arrange(CV.bias, CV.RPMSE, CV.coverage, CV.width, nrow=2))
summary(life.mlr)
CI <- confint(life.mlr, level = 0.95)
CI
# anova(full.lm, reduced.lm)
life.mlr2 <- lm(LifeExp~Group+I(log(PPGDP)), data = life)
anova(life.mlr, life.mlr2)
#End of homework's code

```