Homework 6

Multiple Linear Regression Additional Variable Types

AUTHOR

Zeb Sorenson

Data and Description

Note that for the sake of length for this homework assignment, I am not having you check the model assumptions. You certainly can, if you would like, and in "real life" you would definitely need to do this prior to any statistical inference.

Macroeconomists often speculate that life expectancy is linked with the economic well-being of a country. Macroeconomists also hypothesize that Organisation for Economic Co-operation and Development (OECD) (an international think tank charged with promoting policies that will improve global social and economic well-being) members will have longer life expectancy. To test these hypotheses, the LifeExpectancy.txt data set (found on Canvas) contains the following information:

Variable	Description
LifeExp	Average life expectancy in years
Country	Country name
Group	Is the country a member of OECD, Africa, or other?
PPGDP	Per person GDP (on the log scale)

The Group variable indicates if the country is a member of the OECD, a member of the African continent, or belonging to neither group (other). Note that the Country variable is just for your reference - you will not use this variable in your model.

Download LifeExpectancy.txt, and put it in the same folder as this .qmd file.

0. Replace the text "< PUT YOUR NAME HERE >" (above next to "author:") with your full name.

1. Read in the data set, call it "life", remove the "Row" column, and change the class of any categorical variables to factor variables. Print a summary of the data and make sure the data makes sense.

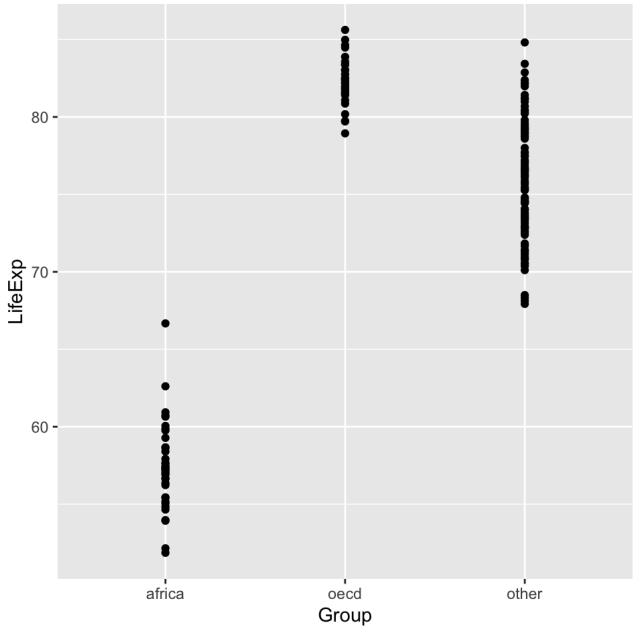
```
life <- read.csv("~/Desktop/Stat 330/LifeExpectancy.txt", sep="")

life <- subset(life, select = -Row)

life<- mutate(life,
              Country = as.factor(Country),
              Group = as.factor(Group))

summary(life)
```

```
     Country       Group         PPGDP           LifeExp     
 Albania  :  1   africa: 37   Min.   : 4.743   Min.   :51.86  
 Anguilla :  1   oecd  : 31   1st Qu.: 7.157   1st Qu.:70.81  
 Argentina:  1   other :113   Median : 8.542   Median :75.29  
 Armenia  :  1                Mean   : 8.492   Mean   :73.14  
 Aruba    :  1                3rd Qu.: 9.844   3rd Qu.:79.80  
 Australia:  1                Max.   :11.563   Max.   :85.62  
 (Other)  :175
```

```
#Everything looks good! The Row column was deleted successfully
```

2. Show a scatterplot with the response on the y-axis and the other continuous variable on the x-axis. Comment on the the relationship between these two variables.

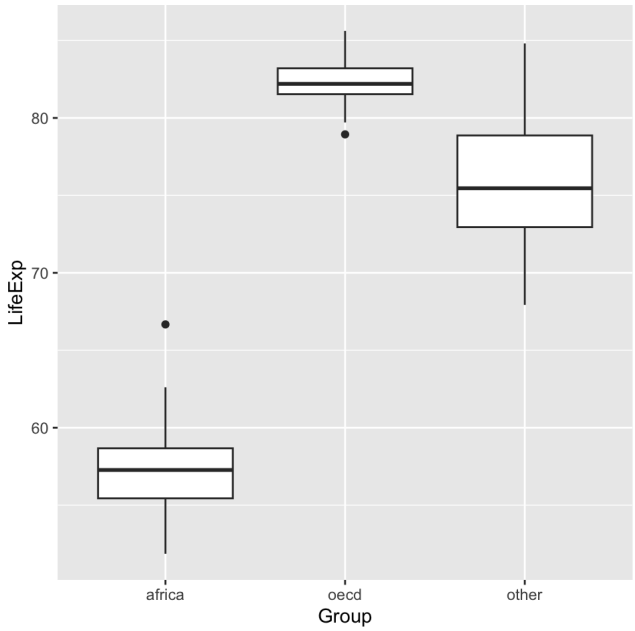
```
ggplot(data = life) +

  geom_point(mapping = aes(x = Group, y = LifeExp)) +

  theme(aspect.ratio = 1)
```



Without making any further analysis at this point, there does appear to be a strong relationship between the Group category and Life Expectancy in years. We can see that the OECD group has the highest life expectancy and smallest range of values. Other is also much higher than Africa but has the widest spread of values.

3. Create and print a boxplot with the response on the y-axis and the categorical variable on the x-axis. Comment on the the relationship between these two variables.

```
ggplot(data = life) +

  geom_boxplot(mapping = aes(x = Group, y = LifeExp)) +

  theme(aspect.ratio = 1)
```



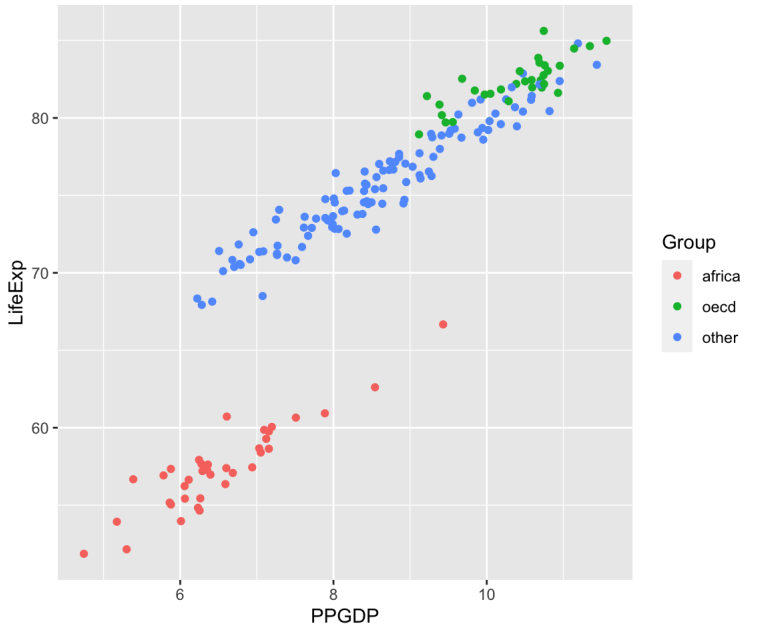
Again, the OECD group has the highest life expectancy with the a median around 83 or 84, which is extremely far from Africa's median which appears to be around 55. The box plot further confirms that the OECD has the

lowest spread of values for life expectancy and while the Other group has a higher life expectancy than Africa, it does also have a wide spread of values for life expectancy in years.

It also appears that both the group of Africa and OECD have outliers. Which may need to be investigated in the future.

4. Create and print a color-coded scatterplot using all of the variables that will be in your model. Hint: plot the response on the y-axis, the other continuous variable on the x-axis, and color the points by the categorical variable.

```
ggplot(data = life) +

  geom_point(mapping = aes(x = PPGDP,
                           y = LifeExp,
                           color = Group)) +
  theme(aspect.ratio = 1)
```



5. Write out the theoretical model (using Greek letters/parameters) that includes main effects for Per Person GDP and the group of the country (you will not write out the fitted model using coefficients, because you have not fit a model yet;)). DO NOT include interactions at this step. Remember, you will need to use dummy variables for Group. **USE "other" AS THE BASELINE CATEGORY**. Use variable names that are descriptive (not y, x_1 , etc.).

$$\text{Life\_Expectancy\_Years}_i = \beta_0 + \beta_1 \times \text{GPD\_PerPerson}_i + \beta_2 \times \text{I}(\text{Group} = \text{Africa}_i) + \beta_3 \times \text{I}(\text{Group} = \text{OECD}_i) + \epsilon_i$$

Where

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

6. Fit the multiple linear regression model from question 5 to the data (no transformations, no interactions, etc.) **using dummy variables that you create manually**. *USE "other" AS THE BASELINE CATEGORY FOR GROUP*. Print a summary of the results.

```
life$Group_OECD <- ifelse(life$Group == "oecd", 1, 0)

life$Group_Africa <- ifelse(life$Group == "africa", 1, 0)

life_lm_overparm <- lm(LifeExp ~
                         PPGDP +
                         Group_OECD +
                         Group_Africa,
                         data = life)
summary(life_lm_overparm )
```

```
Call:
lm(formula = LifeExp ~ PPGDP + Group_OECD + Group_Africa, data = life)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8136 -0.6546 -0.0327  0.6861  3.0454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.95789    0.65272  78.070  < 2e-16 ***
PPGDP          2.87690    0.07478  38.470  < 2e-16 ***
Group_OECD     1.52983    0.25418   6.019 9.88e-09 ***
Group_Africa -12.29427    0.25726 -47.789  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.077 on 177 degrees of freedom
Multiple R-squared:  0.9857,    Adjusted R-squared:  0.9855
F-statistic:  4080 on 3 and 177 DF,  p-value: < 2.2e-16
```

Here we set our linear model with dummy variables that take on a zero or one, since they are categorical and manually insert them into the model. Excluding other means that it is our baseline for this model.

7. Fit the multiple linear regression model from question 5 again, but this time let R create the dummy variables for you in the lm function. As before, *USE "other" AS THE BASELINE CATEGORY FOR GROUP*. Print a summary of the results and make sure they are identical to the results from question 6.

```
levels(life$Group)  #original order of levels
```

```
[1] "africa" "oecd"   "other"
```

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

life_lm_second <- lm(LifeExp ~ Group+ PPGDP, data = life) #Don't include country, we don't care a

summary(life_lm_second)
```

```
Call:
lm(formula = LifeExp ~ Group + PPGDP, data = life)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8136 -0.6546 -0.0327  0.6861  3.0454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.95789    0.65272  78.070  < 2e-16 ***
Groupoecd     1.52983    0.25418   6.019 9.88e-09 ***
Groupafrica -12.29427    0.25726 -47.789  < 2e-16 ***
PPGDP         2.87690    0.07478  38.470  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.077 on 177 degrees of freedom
Multiple R-squared:  0.9857,    Adjusted R-squared:  0.9855
F-statistic:  4080 on 3 and 177 DF,  p-value: < 2.2e-16
```

Here we see that we can either take a manual approach like in question 6, or let R handle the dummy variables for us. Either way, we end up with the same values in our summary. (Even if the order of coefficients is slightly different)

8. Briefly interpret the intercept (like we did in class). **Note that you will need to use the word "average" twice since you are predicting an average already (i.e. the response variable is a country's average life expectancy).** You will need to do this here and with the questions following, when interpreting.

In our model, holding all else constant, when a country's PPGDP is zero and a country belongs to the OTHER category, a country's average life expectancy is on average 50.96 years.

9. Briefly interpret the coefficient for PPGDP. You do not need to un-transform anything or interpret this in the percentage change framework - you can just write something like "for every one unit increase in per person GDP (log scale)" in your response.

For every one unit increase in per person GDP (log scale) the average life expectancy increases by an average of 2.88 years in relation to countries that belong to the the OECD group.

10. For equal per person GDP (log scale), how does life expectancy change for countries that are members of the OECD compared to countries that are on the African continent? Show how you obtained this number, and briefly interpret this number (like we did in class).

For every one unit increase in per person GDP (log scale) the average life expectancy for countries part of the OECD increase by an average of 1.52983, whereas for an equal one unit increase in per person GDP, the average life expectancy for countries part of Africa decrease by an average 12.29427 years.

We can get this number/coefficient by printing out the summary of our model as such,

life_lm_second <- lm(LifeExp ~ Group+ PPGDP, data = life)

11. Create 95% confidence intervals for all coefficients (use the `confint` function). You do not need to interpret them in this question.

```
confint(life_lm_overparm, level = 0.95)
```

```
                 2.5 %     97.5 %
(Intercept)   49.669772  52.246000
PPGDP          2.729321   3.024483
Group_OECD     1.028207   2.031453
Group_Africa -12.801967 -11.786572
```

```
#If we had Other, the intercept would lose meaning,
#The base line IS the intercept that because if a country is not in Africa or OECD, then it falls
```

12. Briefly interpret the 95% confidence interval for I(Group=Africa).

Holding all else constant, we are 95% confident that on average, the average life expectancy for countries belonging to Africa lies between 12.80 and 11.79 years LESS than countries belonging to the Other group.

13. Use the `anova` function to conduct a hypothesis test that tests a reduced model compared to the full model. Specifically, test if Group has a significant effect on LifeExp. What do you conclude from the result of the test? Hint: you will need to create another linear model and compare it with the one you made previously.

```
new_reduced_LM <- lm(LifeExp ~ PPGDP, data=life)

anova(new_reduced_LM, life_lm_overparm)
```

```
Analysis of Variance Table

Model 1: LifeExp ~ PPGDP
Model 2: LifeExp ~ PPGDP + Group_OECD + Group_Africa
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    179 2858.84
2    177  205.48  2    2653.4 1142.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We compare our two models, one with Group included and one without. Our null hypothesis is that there is no significant difference between the two models, meaning that Group does not have a significant effect on life expectancy. However, after running Anova with out two models, we see an extremely small p value of 2.2e-16 meaning that we reject the null and conlude that Group actually does have a significant impact on life expectancy.

14. Create a 95% prediction interval for the life expectancy of a country in the OECD with an average per person GDP (log scale) of 9.5. Print the result, and briefly interpret this interval (like we did in class). (Use the `predict` function.)

```
predict(life_lm_second,

        newdata  = data.frame(Group = 'oecd',
                              PPGDP = 9.5),

        interval = "prediction",

        level = 0.95)
```

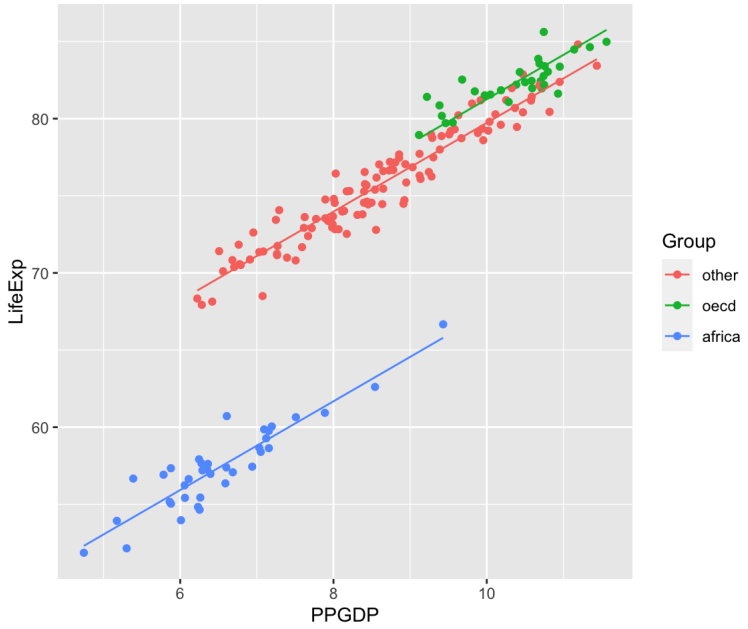
```
       fit      lwr      upr
1 79.81829 77.65424 81.98233
```

```
#Had to use the second LM becasue the first had dummy variables which made it harder to remember
```

Holding all else constant, we are 95% confident that the average life expectancy for a country that is part of the OECD group and has a PPGDP(log scale) of 9.5 will be between 79.82 and 81.98 years.

15. Plot the fitted model on the scatterplot with the two continuous variables on the axes, colored by the categorical variable. Hint: you should have 3 different lines on your plot, and you will *not* need to have different line types or point shapes (you *will* need to have different colors).

```
ggplot(life) +

  geom_point(mapping = aes(x = PPGDP,
                           y = LifeExp,
                           color = Group)) +
  geom_line(mapping = aes(x = PPGDP,
                          y = predict(life_lm_second),
                          color = Group)) +
  theme(aspect.ratio = 1)
```



16. Fit a multiple linear regression model to the data, where this time you **include an interaction term** between PPGDP and Group. *USE "other" AS THE BASELINE CATEGORY FOR GROUP*. Print a summary of the results.

```
levels(life$Group)  # Other should be set as the basline
```

```
[1] "other"  "oecd"   "africa"
```

```
life_lm_interation <- lm(LifeExp ~PPGDP+Group+PPGDP:Group, data=life)

summary(life_lm_interation)
```

```
Call:
lm(formula = LifeExp ~ PPGDP + Group + PPGDP:Group, data = life)

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 ***
PPGDP               2.93882    0.08187  35.896  < 2e-16 ***
Groupoecd          11.29201    3.21337   3.514 0.000562 ***
Groupafrica       -11.89511    1.47535  -8.063 1.13e-13 ***
PPGDP:Groupoecd    -0.95268    0.31279  -3.046 0.002680 **
PPGDP:Groupafrica  -0.04128    0.21249  -0.194 0.846187
```

```
---
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
```

17. Write out the fitted model (using coefficients values from above) for a model with PPGDP, Group, and an interaction between PPGDG and Group. Remember, you will need to use dummy variables for Group. **USE "other" AS THE BASELINE CATEGORY**. Use variable names that are descriptive (not y, x_1, etc.).

$$\widehat{\text{LifeExpectancy}}_i = 50.42 + 2.93882 \cdot (\text{PPGDP}_i) + 11.29201 \cdot \text{I}(\text{Group} = \text{OECD}_i) - 11.89511 \cdot \text{I}(\text{Group} = \text{Africa}_i) - 0.95268 \cdot (\text{PPGI}$$

18. Use the `anova` function to test if the overall interaction between PPGDP and Group is significant. Print the result. What do you conclude (full sentance)?

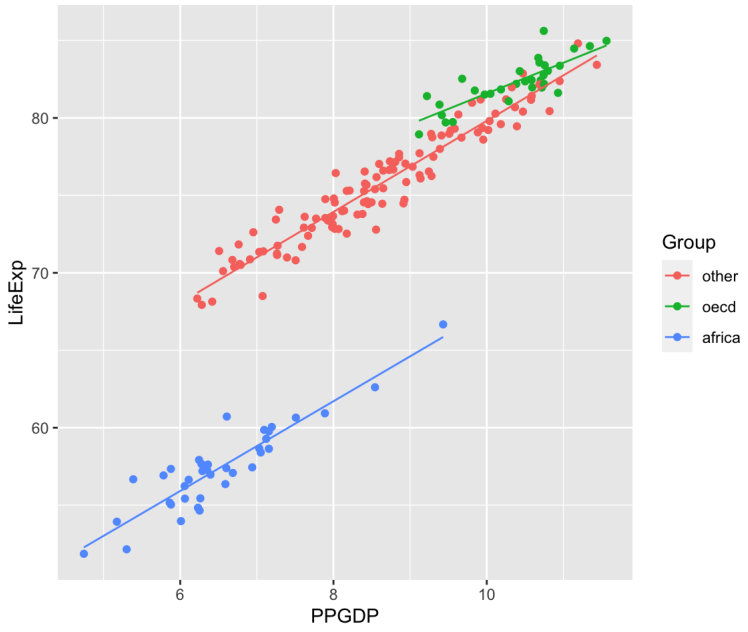
```
anova(life_lm_second,life_lm_interation )
```

```
Analysis of Variance Table

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

We run Anova with a model which includes the interaction of PPGDP and Group and a model that does not include this model. The null hypothesis is that there is no significant difference between the two models. However, our small P value of 0.01083 indicates that we reject the null and conclude that the interaction between PPGDP and Group is indeed significant.

19. Plot the fitted model (with the interaction included) on the scatterplot with the two continuous variables on the axes, colored by the categorical variable. Hint: you should have 3 different lines on your plot, and you will *not* need to have different line types or point shapes (you *will* need to have different colors).

```
ggplot(life) +

  geom_point(mapping = aes(x = PPGDP,
                           y = LifeExp,
                           color = Group)) +
  geom_line(mapping = aes(x = PPGDP,
                          y = predict(life_lm_interation),
                          color = Group)) +
  theme(aspect.ratio = 1)
```



20. How did the fitted lines change when you included an interaction term compared with when you did not include an interaction term?

Comparing this graph to the one we made in question 16, which is the model without the interaction, this model with the interaction is almost identical. Except for the green line for OECD which has shifted towards the x axis (PPGDP). Other than this, everything else looks identical.

21. What is the estimated effect of PPGDP on LifeExp for countries in a country other than those in the OECD or Africa (i.e. in the "other" category)? You should report this number in a complete sentence (as done in class toward the end of the notes). Since this is a continuous-categorical interaction, and since we are focusing on the effect of the continuous variable, you should use the "one unit increase" terminology in your response.

The estimated average effect of PPGD is that for every 1 unit increase of PPGDP (Log Scale) the average life expectancy of countries in the Other category (Not OECD or Africa) will increase by 2.94 years.

We can see this by looking at our model summary in respect for the baseline category (Other).

22. What is the p-value for the test of whether the effect of PPGDP on LifeExp is different between countries in the OECD group and countries in the "other" group?

0.002680. Which we can get from the summary of our fitted model and looking at the corresponding P value to the coefficient PPGDP:Groupoecd.

23. What is the p-value for the test of whether the effect of PPGDP on LifeExp is different between countries in the OECD group and countries in the African continent? (use the glht() function from the multcomp package to get an answer to this question.)

```
# Got help from ChatGPT here...But Joe helped me to understand what is actually going on here :)
contrast_matrix <- matrix(c(0,0,0,0, 1, -1), nrow = 1)
#Matrix which tells R to only look at these two beta values, then taking the literal difference b

linear_hypothesis_test <- glht(life_lm_interation, linfct = contrast_matrix)

summary(linear_hypothesis_test)
```

```
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = LifeExp ~ PPGDP + Group + PPGDP:Group, data = life)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  -0.9114     0.3600  -2.532   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

The P value for question #23 is 0.0122

24. What is the effect of PPGDP on LifeExp for countries in the OECD (relative to the reference group)? You should report a number in a complete sentence (as done in class toward the end of the notes). Since this is a continuous-categorical interaction, and since we are focusing on the effect of the continuous variable, you should use the "one unit increase" terminology in your response.

```
summary(life_lm_interation)
```

```
Call:
lm(formula = LifeExp ~ PPGDP + Group + PPGDP:Group, data = life)

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 ***
PPGDP               2.93882    0.08187  35.896  < 2e-16 ***
Groupoecd          11.29201    3.21337   3.514 0.000562 ***
Groupafrica       -11.89511    1.47535  -8.063 1.13e-13 ***
PPGDP:Groupoecd    -0.95268    0.31279  -3.046 0.002680 **
PPGDP:Groupafrica  -0.04128    0.21249  -0.194 0.846187
```

```
---
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
```

The estimated average effect of PPGD is that for every 1 unit increase of PPGDP (Log Scale), the average life expectancy of countries in the OECD group will decrease by (0.95268+2.94 =3.89268) 3.89 years.

25. Conditional on having a PPGDP of 9, what is the estimated effect of belonging to the OECD relative to being in the "other" country group? You should report a number in a complete sentence (as done in class toward the end of the notes).

```
effect<- 11.29201 +(9*(-0.95268))
effect
```

```
[1] 2.71789
```

When a country/Group has a PPGDP of 9, the estimated effect of belonging to the OECD relative to being in the Other country group is 2.71789

26. Briefly summarize what you learned from this analysis *to a non-statistician*. Write a few sentences about (1) the purpose of this data set and analysis and (2) what you learned about this data set from your analysis. Write your response as if you were addressing a business manager (avoid using statistics jargon) and just provide the main take-aways.

We speculated that life expectancy may be linked to the financial well being of a countries economy. We decided to take this into consideration and also analyze if a country being part of either the OECD or Africa or otherwise not part of these two groups, also played any role in the average life expectancy in that country. Along with this, we also looked at a countries per person GDP to see if it as well had any role in the average like expectancy of a country.

We ran a number of tests to see if any type of relationship could be found with these variables a countries average life expectancy. We were rigorous in our analysis to determine if these variables truly played a role one with another or separately and concluded that both which group a country belonged to and that countries PPGDP did play a significant role in the countries average life expectancy. Ultimately, countries belonging to OECD had the highest life expectancy along the the highest PPGDP. In All cases, higher PPGDP did improve average life expectancy as well.