

Homework 6 | DATA 5600

Multiple Linear Regression Additional Variable Types

Carter Grant

```
# load packages here
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
import statsmodels.api as sm
from statsmodels.graphics.factorplots import interaction_plot
from statsmodels.stats.anova import anova_lm
```

✓ 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 this Jupyter Notebook.

0. Replace the text "< YOUR NAME HERE >" (above) with your full name.

- ✓ 1. Read in the data set, call it "life" and remove the "Row" column. Print a summary of the data, and expore the data to make sure the data makes sense. [1 point]

```
life = pd.read_table("/content/LifeExpectancy.txt", delimiter = ' ')
life.head()
```




	Row	Country	Group	PPGDP	LifeExp
0	1	Albania	other	8.209907	75.31
1	2	Anguilla	other	9.528801	79.19
2	3	Argentina	other	9.122831	77.72
3	4	Armenia	other	8.016549	74.54
4	5	Aruba	other	10.036772	79.80

```
life.drop("Row", axis = 1, inplace = True)
life.head()
```



	Country	Group	PPGDP	LifeExp
0	Albania	other	8.209907	75.31
1	Anguilla	other	9.528801	79.19
2	Argentina	other	9.122831	77.72
3	Armenia	other	8.016549	74.54
4	Aruba	other	10.036772	79.80

```
print(life.info())
```




<class 'pandas.core.frame.DataFrame'>				
RangeIndex: 181 entries, 0 to 180				
Data columns (total 4 columns):				
#	Column	Non-Null Count	Dtype	
0	Country	181 non-null	object	
1	Group	181 non-null	object	
2	PPGDP	181 non-null	float64	
3	LifeExp	181 non-null	float64	
dtypes: float64(2), object(2)				
memory usage: 5.8+ KB				
None				

```
life['Group'].value_counts()
life['Country'].value_counts()
```



	count
Country	
Albania	1
Netherlands	1
NewZealand	1
Nicaragua	1
Niger	1
...	...
Ghana	1
Greece	1
Greenland	1
Grenada	1
Zimbabwe	1
181 rows × 1 columns	
dtype: int64	

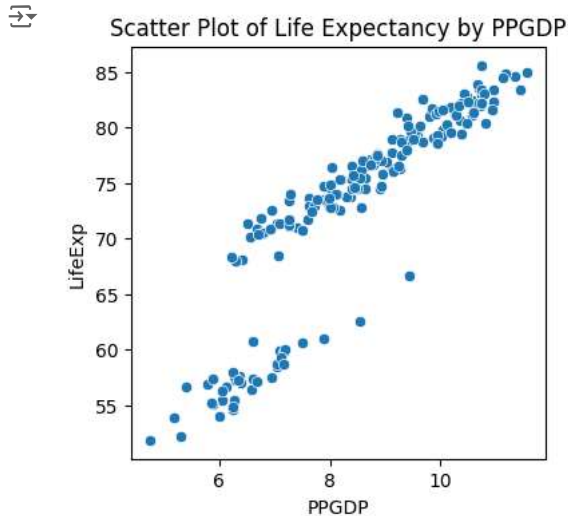
```
life['PPGDP'].describe()
life['LifeExp'].describe()
```



	LifeExp
count	181.000000
mean	73.136630
std	8.948747
min	51.860000
25%	70.810000
50%	75.290000
75%	79.800000
max	85.620000
dtype: float64	

- ✓ 2. Create and print 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. [2 points]

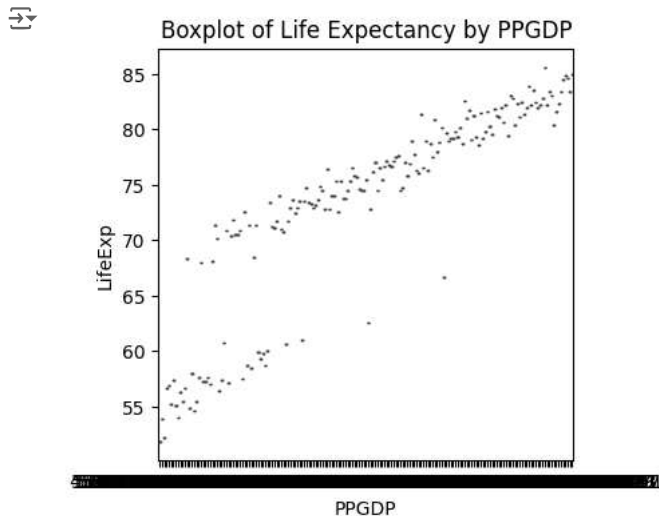
```
plt.figure(figsize=(4, 4)) # Create a new figure for the scatter plot
sns.scatterplot(data=life, x='PPGDP', y='LifeExp')
plt.title('Scatter Plot of Life Expectancy by PPGDP')
plt.show()
```



It seems that when Life Expectancy increases, PPGDP will also increase as seen in the graph's trend above.

- ✓ 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. [2 points]

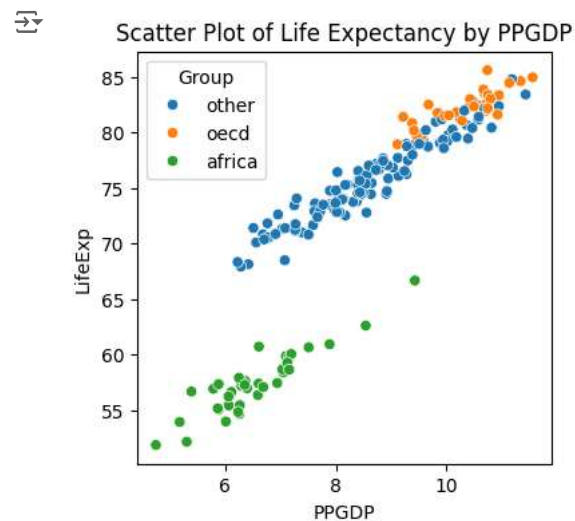
```
plt.figure(figsize=(4, 4))
sns.boxplot(x='PPGDP', y='LifeExp', data=life)
plt.xlabel('PPGDP')
plt.ylabel('LifeExp')
plt.title('Boxplot of Life Expectancy by PPGDP')
plt.show()
```



When Life Expectancy increases, PPGDP will also increase as seen in the graph's trend above. (Note: I am not sure why it is outputting a scatterplot as my code is for boxplot)

- ✓ 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. [1 point]

```
plt.figure(figsize=(4, 4)) # Create a new figure for the scatter plot
sns.scatterplot(data=life, x='PPGDP', y='LifeExp', hue='Group')
plt.title('Scatter Plot of Life Expectancy by PPGDP')
plt.show()
```



5. Write out the general/theoretical model (using Greek letters/parameters) that you are thinking about applying to this data set (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.). [2 points]

$LifeExp_i = \beta_0 + \beta_1 PPGDP_i + \beta_2 Group_{i,OECD} + \beta_3 Group_{i,Africa} + \epsilon_i$

(note: the i is supposed to be subscript i for each variable that is a prediction)

6. Code indicator/dummy variables for Group. [1 point]

```
life_dummy = pd.get_dummies(life,
                             columns = ['Group'],
                             dtype='int')
life_dummy.head()
```

	Country	PPGDP	LifeExp	Group_africa	Group_oecd	Group_other
0	Albania	8.209907	75.31	0	0	1
1	Anguilla	9.528801	79.19	0	0	1
2	Argentina	9.122831	77.72	0	0	1
3	Armenia	8.016549	74.54	0	0	1
4	Aruba	10.036772	79.80	0	0	1

7. Fit a multiple linear regression model to the data (no transformations, no interactions, etc.) using "other" as the baseline for group. Print a summary of the results. [1 point]

```
y = life_dummy['LifeExp']
X_full = sm.add_constant(life_dummy[['PPGDP', 'Group_africa', 'Group_oecd']])
mod_full = sm.OLS(y, X_full)
res_full = mod_full.fit()
res_full.summary()
```



OLS Regression Results

Dep. Variable: LifeExp **R-squared:** 0.986
Model: OLS **Adj. R-squared:** 0.986
Method: Least Squares **F-statistic:** 4080.
Date: Tue, 29 Oct 2024 **Prob (F-statistic):** 4.48e-163
Time: 01:10:32 **Log-Likelihood:** -268.31
No. Observations: 181 **AIC:** 544.6
Df Residuals: 177 **BIC:** 557.4
Df Model: 3

Covariance Type: nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	50.9579	0.653	78.070	0.000	49.670	52.246
PPGDP	2.8769	0.075	38.470	0.000	2.729	3.024
Group_africa	-12.2943	0.257	-47.789	0.000	-12.802	-11.787
Group_oecd	1.5298	0.254	6.019	0.000	1.028	2.031
Omnibus:	0.441	Durbin-Watson:	2.080			
Prob(Omnibus):	0.802	Jarque-Bera (JB):	0.333			
Skew:	0.105	Prob(JB):	0.847			
Kurtosis:	3.015	Cond. No.	75.0			

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

8. Briefly interpret the intercept (like we did in class). **Note that you will need to use the word "average" (or similar) twice since you are predicting an average already.** You will need to do this (say average twice) here and with the questions that follow when interpreting. [2 points]

The intercept of the regression model is approximately 50.96, indicating that this is the predicted average life expectancy for countries in the "other" group, when the Per Capita GDP (PPGDP) is 0. This means that when a country has an average PPGDP of zero, the average life expectancy would be about 50.96 years.

9. Briefly interpret the coefficient for PPGDP (log scale) (like we did in class). You do not need to un-transform anything - you can just write something like "per person GDP (log scale)" in your response. [2 points]

The coefficient for PPGDP is approximately 2.88, which indicates that a 1% increase in per person GDP (on a log scale) is associated with an expected increase in average life expectancy of 2.88 years.

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). [2 points]

Difference = $1.53 - (-12.29) = 13.82$ (this is from the coefficients in the ols test above)

This result of 13.82 indicates that, for countries with equal per person GDP, those that are members of the OECD have an average life expectancy that is about 13.82 years higher than that of countries located on the African continent.

11. Briefly interpret the 95% confidence interval for I(Group=Africa). [2 points]

```
confidence_intervals = res_full.conf_int()
group_africa_confidence_interval = confidence_intervals.loc['Group_africa']
print("95% Confidence Interval for Group_Africa:", group_africa_confidence_interval.values)
```



Show hidden output

The coefficient for Group_Africa is approximately -12.29, with a 95% confidence interval of (-12.80, -11.79). This confidence interval means that we can be 95% confident that the true effect of being in the African group, compared to the baseline group "other," results in an average life expectancy that is between 11.79 and 12.80 years lower (holding all else constant).

- ✓ 12. Use the `anova_lm` function from `statsmodels` to conduct a hypothesis test that tests if Group as a whole 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. [2 points]

```
reduced_model = sm.OLS(y, sm.add_constant(life_dummy[['PPGDP']])).fit()
anova_lm(reduced_model, res_full)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	179.0	2858.842765	0.0	NaN	NaN	NaN
1	177.0	205.477622	2.0	2653.365143	1142.814545	6.409470e-102

Given the extremely low p-value (basically 0), we reject the null hypothesis. This indicates that the Group variable as a whole has a statistically significant effect on average life expectancy (LifeExp). Accounting for the differences in the Group variable (e.g., OECD vs. Africa) significantly improves the model's ability to predict life expectancy when controlling for per person GDP (PPGDP).

13. Create a 95% prediction interval for the average 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 `get_prediction` function on the OLS object.) [2 points]

```
pred = res_full.get_prediction([1, 9.5, 1, 0])

# Get the summary frame with a 95% prediction interval
pred_summary = pred.summary_frame(alpha=0.05).iloc[:, [0, 4, 5]]

print(pred_summary)
```

	mean	obs_ci_lower	obs_ci_upper
0	65.994187	63.795156	68.193217

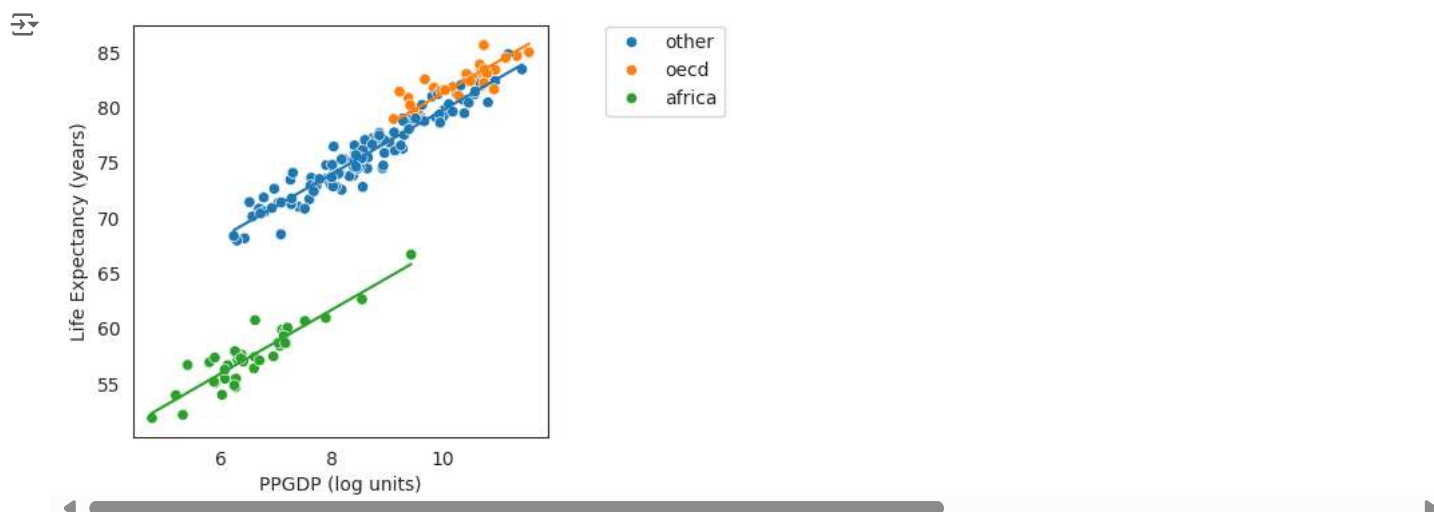
The 95% prediction interval for the average life expectancy of a country in the OECD with an average per person GDP (log scale) of 9.5 is between 63.795 and 68.193 years. This means we are 95% confident that the average life expectancy for such a country will fall within the range (63.795, 68.193)

14. 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). [1 point]

```
print(life_dummy.columns)
```

➡ Show hidden output

```
sns.set_style("white")
plt.figure(figsize = (4, 4))
sns.scatterplot(data = life,
                x = 'PPGDP',
                y = 'LifeExp',
                hue = 'Group')
sns.lineplot(x = life['PPGDP'],
             y = res_full.fittedvalues,
             hue = life['Group'],
             legend = False)
plt.xlabel('PPGDP (log units)')
plt.ylabel('Life Expectancy (years)')
plt.legend(loc = 'upper right', bbox_to_anchor = (1.45, 1.02))
plt.show()
```



15. Fit a multiple linear regression model to the data **using the dummy variables you created**, and include an interaction term between PPGDP and Group. *USE "other" AS THE BASELINE CATEGORY FOR GROUP.* Print a summary of the results. [1 point]

```
life_dummy = pd.get_dummies(life, columns=['Group'], drop_first=False)

life_dummy['PPGDP'] = pd.to_numeric(life_dummy['PPGDP'], errors='coerce')
life_dummy['LifeExp'] = pd.to_numeric(life_dummy['LifeExp'], errors='coerce')

# Convert dummy variables to integers
life_dummy['Group_oecd'] = life_dummy['Group_oecd'].astype(int)
life_dummy['Group_africa'] = life_dummy['Group_africa'].astype(int)
life_dummy['Group_other'] = life_dummy['Group_other'].astype(int)

# Check for any NaN values after conversion
if life_dummy[['PPGDP', 'LifeExp']].isnull().any().any():
    print("NaN values found in 'PPGDP' or 'LifeExp'. Dropping rows with NaN values.")
    life_dummy.dropna(subset=['PPGDP', 'LifeExp'], inplace=True)

# Create interaction terms
life_dummy['PPGDP_Group_oecd'] = life_dummy['PPGDP'] * life_dummy['Group_oecd']
life_dummy['PPGDP_Group_africa'] = life_dummy['PPGDP'] * life_dummy['Group_africa']

X = life_dummy[['PPGDP', 'Group_oecd', 'Group_africa', 'PPGDP_Group_oecd', 'PPGDP_Group_africa']]
y = life_dummy['LifeExp']
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())
```

```
OLS Regression Results
=====
Dep. Variable: LifeExp R-squared: 0.986
Model: OLS Adj. R-squared: 0.986
Method: Least Squares F-statistic: 2551.
Date: Tue, 29 Oct 2024 Prob (F-statistic): 1.98e-161
Time: 01:10:34 Log-Likelihood: -263.63
No. Observations: 181 AIC: 539.3
Df Residuals: 175 BIC: 558.4
Df Model: 5
Covariance Type: nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	50.4240	0.713	70.734	0.000	49.017	51.831
PPGDP	2.9388	0.082	35.896	0.000	2.777	3.100
Group_oecd	11.2920	3.213	3.514	0.001	4.950	17.634
Group_africa	-11.8951	1.475	-8.063	0.000	-14.807	-8.983
PPGDP_Group_oecd	-0.9527	0.313	-3.046	0.003	-1.570	-0.335
PPGDP_Group_africa	-0.0413	0.212	-0.194	0.846	-0.461	0.378

```
=====
Omnibus: 0.578 Durbin-Watson: 2.155
Prob(Omnibus): 0.749 Jarque-Bera (JB): 0.381
```

Skew:	0.105	Prob(JB):	0.827
Kurtosis:	3.078	Cond. No.	374.

=====

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

16. Write out the fitted model (using coefficients values from above) for a model with PPGDP, Group, and an interaction between PPGDP 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.). [3 points]

LifeExp_i = 50.4240 + 2.9388 · PPGDP_i + 11.2920 · Group_{oecd}_i - 11.8951 · Group_{africa}_i - 0.9527 · (PPGDP_i · Group_{oecd}_i) - 0.0413 · (PPGDP_i · Group_{africa}_i) + ϵ

17. Use the `anova_lm` function from `statsmodels` to test if the overall interaction between PPGDP and Group is significant. Print the result. What do you conclude? [2 points]

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

life_dummy = pd.get_dummies(life, columns=['Group'], drop_first=False)

life_dummy['PPGDP'] = pd.to_numeric(life_dummy['PPGDP'], errors='coerce')
life_dummy['LifeExp'] = pd.to_numeric(life_dummy['LifeExp'], errors='coerce')

life_dummy.dropna(subset=['PPGDP', 'LifeExp'], inplace=True)

# Create interaction terms
life_dummy['PPGDP_Group_oecd'] = life_dummy['PPGDP'] * life_dummy['Group_oecd']
life_dummy['PPGDP_Group_africa'] = life_dummy['PPGDP'] * life_dummy['Group_africa']

full_model = ols('LifeExp ~ PPGDP + Group_oecd + Group_africa + PPGDP_Group_oecd + PPGDP_Group_africa', data=life_dummy).fit()

reduced_model = ols('LifeExp ~ PPGDP + Group_oecd + Group_africa', data=life_dummy).fit()

anova_results = anova_lm(reduced_model, full_model)

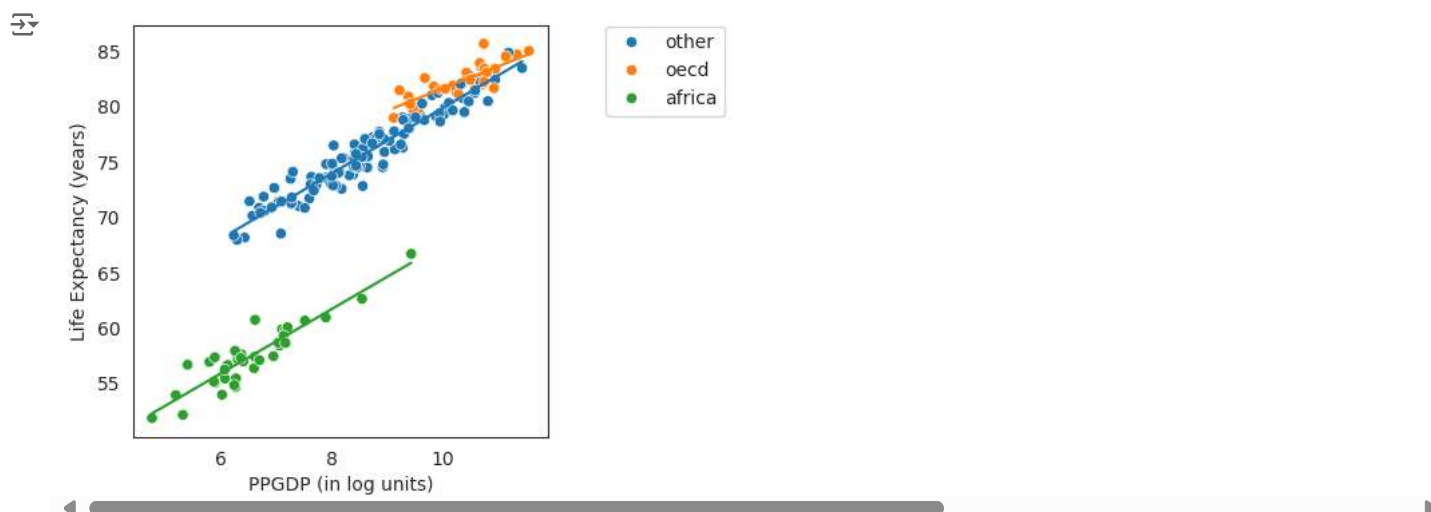
print(anova_results)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	177.0	205.477622	0.0	NaN	NaN	NaN
1	175.0	195.120227	2.0	10.357395	4.644685	0.010828

The p-value (0.010828) is less than 0.05, which indicates that the interaction between PPGDP and Group is statistically significant. This suggests that the effect of PPGDP on life expectancy varies significantly across the different groups (OECD and Africa).

18. 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). [1 point]

```
plt.figure(figsize = (4, 4))
sns.scatterplot(x = life['PPGDP'],
                y = life['LifeExp'],
                hue = life['Group'])
sns.lineplot(x = life['PPGDP'],
              y = full_model.fittedvalues,
              hue = life['Group'],
              legend = False)
plt.xlabel('PPGDP (in log units)')
plt.ylabel('Life Expectancy (years)')
plt.legend(loc = 'upper right', bbox_to_anchor = (1.45, 1.02))
plt.show()
```

19. How did the fitted lines change when you included an interaction term compared with when you did not include an interaction term? [1 point]

The lines seem to be more accurate to fitting the data points. Comparing this plot to the first one, I can see that the lines are closer to the center of the data points.

20. What is the 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 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. [2 points]

The effect of PPGDP on life expectancy for countries in the "other" category (the baseline group) is approximately 2.94 years. This means that for each one unit increase in PPGDP (log units), the life expectancy for countries classified as "other" increases by about 2.94 years, holding all other factors constant.

21. What is the effect of PPGDP on LifeExp for countries in the OECD? 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. [2 points]

For countries in the OECD, the effect of PPGDP on life expectancy is approximately 1.99 years. This means that for each one unit increase in PPGDP (log units), the life expectancy for countries in the OECD increases about 1.99 years, holding all other factors constant.

22. What is the effect of PPGDP on LifeExp for countries in Africa? 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. [2 points]

For countries in Africa, the effect of PPGDP on life expectancy is approximately 2.90 years. This means that for each one unit increase in PPGDP (log units), the life expectancy for countries in Africa is increased about 2.90 years, holding all other factors constant.

23. What is the effect of belonging to the OECD on LifeExp for countries with a PPGDP of 9? You should report a number in a complete sentence (as done in class toward the end of the notes). [2 points]

$\#LifeExp_i = 50.4240 + 2.9388 \cdot 9 + 11.2920 - 0.9527 \cdot 9 + \epsilon_i$

For countries in the OECD with a PPGDP of 9, the expected average life expectancy is approximately 79.59 years, which includes an increase of about 11.29 years compared to countries in the baseline category ("other").

- ✓ 24. What is the effect of belonging to the OECD on LifeExp for countries with a PPGDP of 11? You should report a number in a complete sentence (as done in class toward the end of the notes). [2 points]

#LifeExpi=50.4240+2.9388·11+11.2920·0.9527·11+ei

For countries in the OECD with a PPGDP of 11, the average expected life expectancy is approximately 82.03 years, which includes an increase of about 11.29 years compared to countries in the baseline category ("other").

25. 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. [1 point]
- ✓

The analysis helped us to understand how per capita GDP (PPGDP) impacts life expectancy across different groups of countries, including those in the OECD, Africa, and other regions. It showed that higher PPGDP is generally linked to increased life expectancy, with OECD countries experiencing a significant advantage in average life expectancy compared to others. This emphasizes that economic growth not only enhances GDP but also contributes to better health outcomes, suggesting that policies focused on boosting GDP could lead to public health benefits.