Final Project

I agree to abide by the Stern Code of Conduct - Hugo Pinto, hcp273

5/8/2019

Introduction

OUTCOMES AND PREDICTORS

- In the analysis, the two outcomes are “forestarea” and “airpollution” which are Forest area (% of land area) and PM2.5 air pollution, mean annual exposure (micrograms per cubic meter). The three predictors used are “CO2”, “renewableenergy,” and “popgrowth” which are CO2 emissions (metric tons per capita), renewable energy consumption (% of total final energy consumption), and population growth (annual %).

FOCUS & RELATIONSHIP

- The predictor which will be the focus is population growth, “popgrowth,” to attempt an understanding of how population growth relates to the environment. I assume there would be a positive correlational relationship between pollution and population growth. Additionally, I assume a negative correlational relationship between the forest area and population growth. Both are indicating a deteriorating environment as the population grows.

Analysis

Data description

Units of measure

- airpollution unit of measurement is mean annual exposure (micrograms per cubic meter)
- forestarea unit of measurement is % of land area
- CO2 emissions unit of measurement is metric tons per capita
- renewableenergy consumption unit of measurement is % of total final energy consumption
- popgrowth (population growth) unit of measurement is annual %

first outcome

complex model

```
complex1 <- lm(airpollution ~ CO2 + renewableenergy + popgrowth, wb_recent)

summary(complex1)
```

```
##
## Call:
## lm(formula = airpollution ~ CO2 + renewableenergy + popgrowth,
##     data = wb_recent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.150 -11.676  -4.514   8.876  73.389
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     24.539710   3.028555   8.103 9.11e-14 ***
## CO2             -0.202212   0.258053  -0.784    0.434
## renewableenergy  0.005296   0.062858   0.084    0.933
## popgrowth        4.874299   1.057717   4.608 7.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.3 on 174 degrees of freedom
## Multiple R-squared:  0.1257, Adjusted R-squared:  0.1106
## F-statistic: 8.337 on 3 and 174 DF,  p-value: 3.275e-05
```

MODEL1 - (complex1) Interpretation

- The Coefficient estimate for popgrowth suggests that 5.57805 is the expected change in airpollution (micrograms per cubic meter) if we increase the main predictor, popgrowth, by one of its unit, annual percentage, while holding all the other predictor variables constant.
- The P-value tests the null hypothesis that the coefficient (popgrowth) is equal to zero. The p-value is below the significant level (< 0.05), indicating we reject the null hypothesis. There is statistical significance.
- R-squared is a measure of the data's closeness to the fitted regression line. The higher the R-squared is, the better the model is as a fit of the provided data. The R-squared for this model is 0.15, suggesting, but not certain, that it is not a good fit.
- Adjusted R-squared: R-squared does not imply any information in terms of bias, which makes assessing the residual plots important. Moreover, the more predictors, the larger the R-square. As a result, a model with more predictors may appear to have a better fit simply due to the number of predictors. The adjusted R-squared penalizes for the additional predictors. Similarly, the adjusted R-squared is too low, at 0.1354, for an implied, but not certain, good fit.

sub model

```
sub1 <- lm(airpollution ~ popgrowth + CO2, wb_recent)

summary(sub1)
```

```
##
## Call:
## lm(formula = airpollution ~ popgrowth + CO2, data = wb_recent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.121 -11.593  -4.435   8.776  73.660
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.7204     2.1328  11.590  < 2e-16 ***
## popgrowth     4.9040     0.9946   4.931 1.89e-06 ***
## CO2          -0.2144     0.2133  -1.005    0.316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.25 on 175 degrees of freedom
## Multiple R-squared:  0.1256, Adjusted R-squared:  0.1157
## F-statistic: 12.57 on 2 and 175 DF,  p-value: 7.901e-06
```

MODEL1 - (sub1) Interpretation

- The sub1 model holds similar final interpretations of the complex model.
- The Coefficient estimate for popgrowth suggests that 5.5120 is the expected change in airpollution (micrograms per cubic meter) if we increase the main predictor, popgrowth, by one of its unit, annual percentage, while holding all the other predictor variables constant.
- The P-value tests the null hypothesis that the coefficient (popgrowth) is equal to zero. The p-value is below the significant level (< 0.05), indicating we reject the null hypothesis. There is statistical significance.
- R-squared is a measure of the data's closeness to the fitted regression line. The higher the R-squared is, the better the model is as a fit of the provided data. The R-squared for this model is 0.1498, suggesting, but not certain, that it is not a good fit.
- adjusted R-squared: R-squared does not imply any information in terms of bias, which makes assessing the residual plots important. Moreover, the more predictors, the larger the R-square. As a result, a model with more predictors may appear to have a better fit simply due to the number of predictors. The adjusted R-squared penalizes for the additional predictors. Similarly, the adjusted R-squared is too low, at 0.1402, for an implied, but not certain, good fit.

F-test

```
anova(complex1, sub1)
```

```
## Analysis of Variance Table
##
## Model 1: airpollution ~ CO2 + renewableenergy + popgrowth
## Model 2: airpollution ~ popgrowth + CO2
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    174 52087
## 2    175 52089 -1   -2.1248 0.0071  0.933
```

Which model does the test choose? sub1 model

- The P-value is higher than the 5% significant level, so we fail to reject the null that the simple model is good enough. Sub1, the simple, model, is better after doing this F-test.

complex model confidence intervals

```
#show confints for complex model
confint(complex1, "popgrowth")
```

```
##              2.5 %   97.5 %
## popgrowth 2.786693 6.961905
```

- When predicting “airpollution” with the main predictor, if we do the same procedure for numerous times, for around 95% of all the time the confidence intervals of “popgrowth” is, at its lower bound, 3.449114 and 7.706985 at the upper bound in the model. There is a statistical significance because it does not contain the null hypothesis value, 0, and which may be taken practically because as we can see as the population grows, potentially so will air pollution which applies to the real world. As the population increases, we grow higher demand for polluting resources. Currently, the U.N promotes a decrease or maintenance of pollution. Pollution increase is not only harmful to our environment but also exposed humans.

sub model confidence intervals

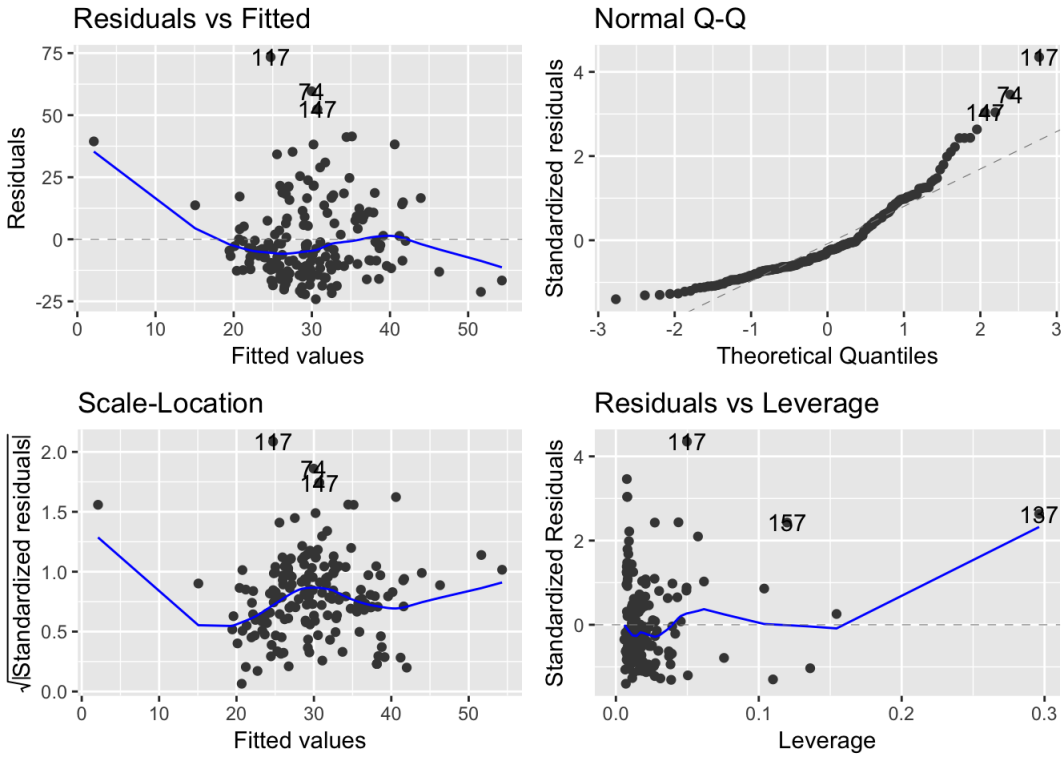
```
#show confints for the sub model
confint(sub1, "popgrowth")
```

```
##              2.5 %   97.5 %
## popgrowth 2.941075 6.866849
```

- The submodel1 also suggests there is a statistical significance because it does not contain the null hypothesis value, 0, and which may be taken practically because as we can see as the population grows, potentially so will air pollution which applies to the real world. This model when compared to the complex, has a small width in the confidence intervals.

complex1 model diagnostic plots

```
#diagnostic plots
autoplot(complex1)
```



Residuals vs. Fitted (top left)

- This plot shows a pattern that the values mostly are located in a specific area, indicating possible non-linear relationships that the linear model has not captured. The vertical spread of the points is not the same throughout the whole graph, mostly in one area, looking like a triangle pointing slightly left, indicating heteroscedastic errors.

Normal Q-Q (top right)

- In this plot, it demonstrates that the residuals are not normally distributed because there are points that fall outside of the dashed line.

Scale-Location (bottom left)

- Similarly to the Residuals vs. Fitted Plot, this plot indicates possible non-linear relationships that the linear model has not captured. Also shows heteroscedastic errors.

Residuals vs. Leverage (bottom right)

- In this plot, there lies a point to the far right, its residual slightly far from zero, is a point of high leverage and comparably an outlier. Suggesting a bad fit.

ASSUMPTIONS:

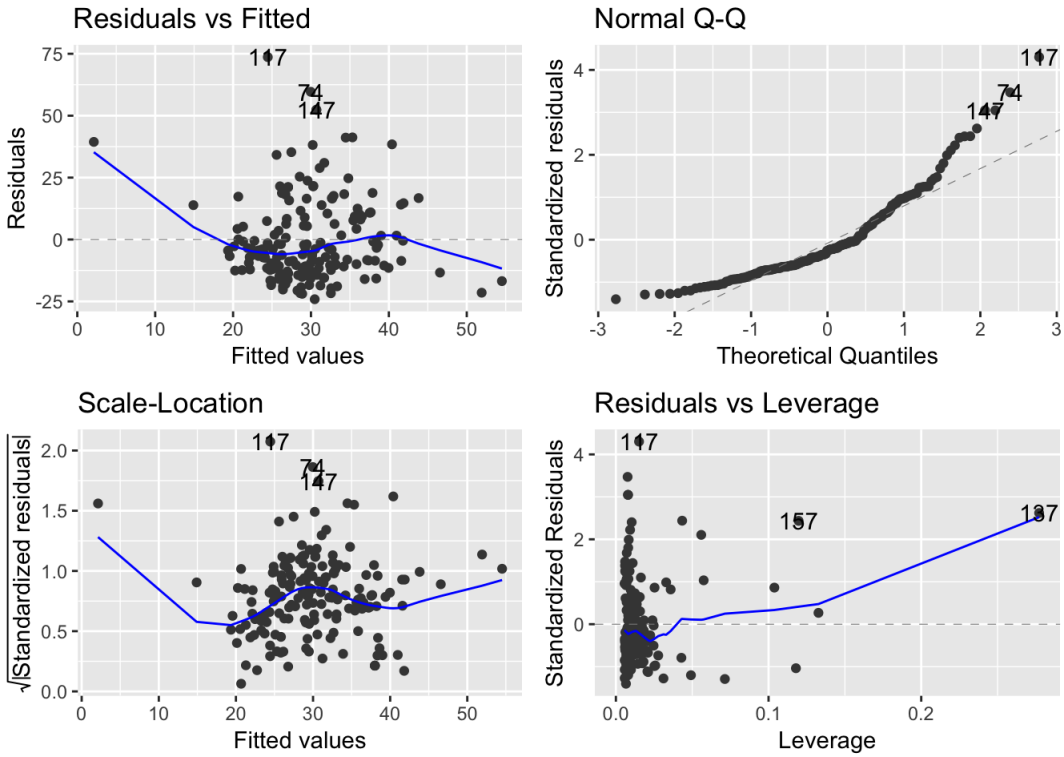
- According to the textbook, OpenIntroStatistics, Multiple regression model generally depend on the following four assumptions: the residuals of the model are nearly normal, the variability of the residuals is nearly constant, the residuals are independent, and each variable is linearly related to the outcome.

CONCLUDE:

- The diagnostics plots do not support the model assumptions, and this lessens the credibility of the model. The plots show many shortcomings stated in the plot's interpretations.

submodel1 diagnostic plots

```
autoplot(sub1)
```



CONCLUDE:

- Interpreting the submodel diagnostic plots suggests it holds the same problems seen on the complex model plots with no key differences. The expected positive correlational relationship between the outcome and the main predictor could be seen from the F-test chosen model, sub1, confidence interval of 3.53309 and 7.490961, and this suggested the presence of statistical and practical significance. Unfortunately, however, neither model diagnostic plots suffice the multiple regression model assumptions. Therefore, the complex1 model nor the sub1 model are not good enough due to their shortcomings, indicating that the positive correlational relationship might not be credible enough.

Second outcome

```
#complex2 model
complex2 <- lm(forestarea ~ CO2 + renewableenergy + popgrowth, wb_recent)

summary(complex2)
```

```
##
## Call:
## lm(formula = forestarea ~ CO2 + renewableenergy + popgrowth,
##     data = wb_recent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.262 -13.386  -1.383  10.734  69.212
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     25.94901    3.33315   7.785 5.97e-13 ***
## CO2              0.35296    0.28401   1.243 0.215618
## renewableenergy  0.26997    0.06918   3.902 0.000136 ***
## popgrowth       -4.57889    1.16409  -3.933 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.04 on 174 degrees of freedom
## Multiple R-squared:  0.1232, Adjusted R-squared:  0.108
## F-statistic: 8.146 on 3 and 174 DF,  p-value: 4.171e-05
```

```
#sub2 model
sub2 <- lm(forestarea ~ CO2 + popgrowth, wb_recent)

summary(sub2)
```

```
##
## Call:
## lm(formula = forestarea ~ CO2 + popgrowth, data = wb_recent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.959 -17.235  -0.320   9.981  67.443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.1576     2.4479  14.363  < 2e-16 ***
## CO2          -0.2668     0.2448  -1.090  0.27728
## popgrowth    -3.0668     1.1415  -2.687  0.00791 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.8 on 175 degrees of freedom
## Multiple R-squared:  0.04641,    Adjusted R-squared:  0.03551
## F-statistic: 4.259 on 2 and 175 DF,  p-value: 0.01563
```

MODEL2 - (complex2) Interpretation

- The Coefficient estimate for popgrowth in the complex2 model suggests that -5.1086 is the expected change in forestarea (% of land area) if we increase the main predictor, popgrowth, by one of its unit, annual percentage, while holding all the other predictor variables constant. Additionally, for the sub2 model -3.3467 is the expected change in forestarea (% of land area) if we increase the main predictor, popgrowth, by one of its unit, annual percentage, while holding all the other predictor variables constant. Notice there is a smaller change predicted in the submodel, however still negative.
- The P-value tests the null hypothesis that the coefficient (popgrowth) is equal to zero. The p-value for the main predictor in both, complex2 and submodel2, are below the significant level (< 0.05) indicating we reject the null hypothesis. There is statistical significance.
- R-squared is a measure of the data's closeness to the fitted regression line. The higher the R-squared is, the better the model is as a fit of the provided data. The R-squared for complex2 model is 0.1336, suggesting, but not certain, that it is not a good fit. Additionally, the R-squared value for the sub2 model is smaller, 0.05188, but also suggests a not good fit.
- adjusted R-squared: R-squared does not imply any information in terms of bias, which makes assessing the residual plots important. Moreover, the more predictors, the larger the R-square. As a result, a model with more predictors may appear to have a better fit simply due to the number of predictors. The adjusted R-squared penalizes for the additional predictors. Similarly, the adjusted R-squared is too low for the complex2 and sub2 models respectively, at 0.1187 and 0.04111, for an implied, but not certain, good fit.

F-test

```
anova(complex2, sub2)
```

```
## Analysis of Variance Table
##
## Model 1: forestarea ~ CO2 + renewableenergy + popgrowth
## Model 2: forestarea ~ CO2 + popgrowth
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    174 63091
## 2    175 68613 -1   -5521.7 15.229 0.000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which model does the test choose? complex2

- The P-value is lower than the 5% significant level, so we reject the null that the simple model is good enough. The more complex model, complex2, is better than the sub2 model after doing this F-test.

complex2 and sub2 confidence intervals

```
#show confints
confint(complex2, "popgrowth")
```

```
##               2.5 %    97.5 %
## popgrowth -6.876457 -2.281329
```

```
#show confints
confint(sub2, "popgrowth")
```

```
##                2.5 %     97.5 %
## popgrowth -5.319576 -0.8139614
```

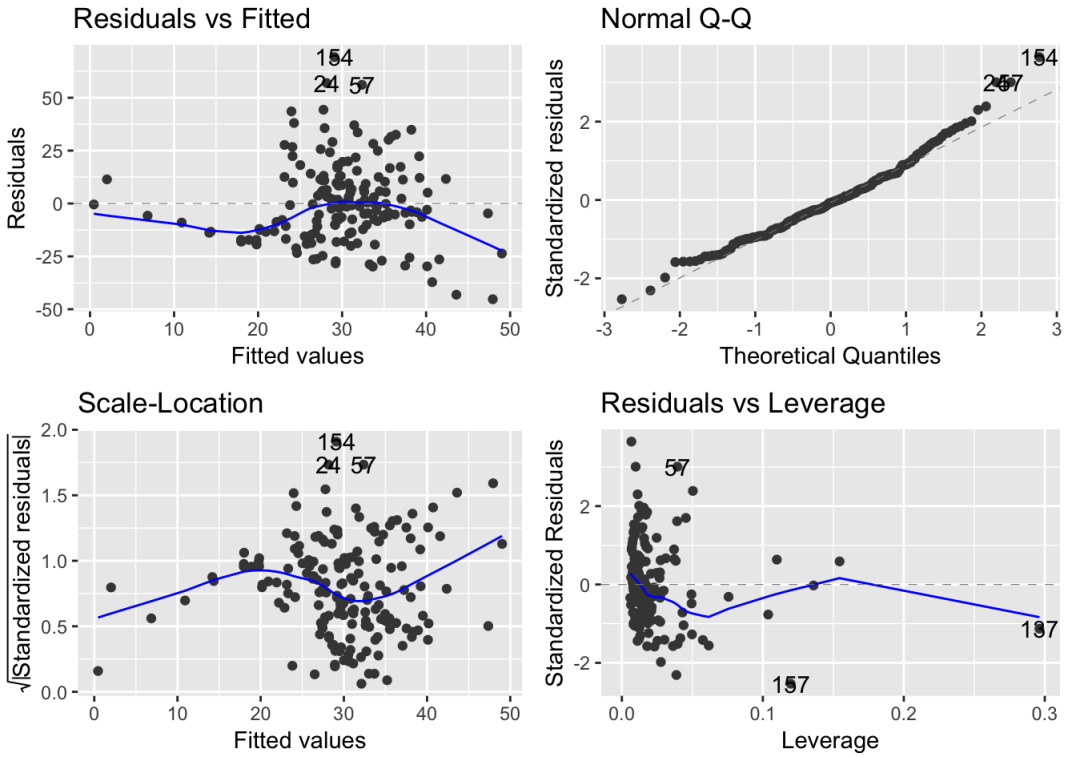
- When predicting “forestarea” in complex2 with the main predictor, if we do the same procedure for numerous times, for around 95% of all the time the confidence interval of “popgrowth” is, at its lower bound, -7.473069 and, at its upper bound, -2.744234 in the model. In the sub2 model, the confidence interval is at its lower bound, -5.645635 and, at upper bound, -1.047797. For both, there is a statistical significance because it does not contain the null hypothesis value, 0, and which may be taken practically because as we can see as the population grows, potentially decreasing forest area which applies to the real world. As population increases, we grow higher demand for land space and agriculture, in turn, deforestation, at any decreasing rate being harmful to the environment.

complex and sub models diagnostic plots

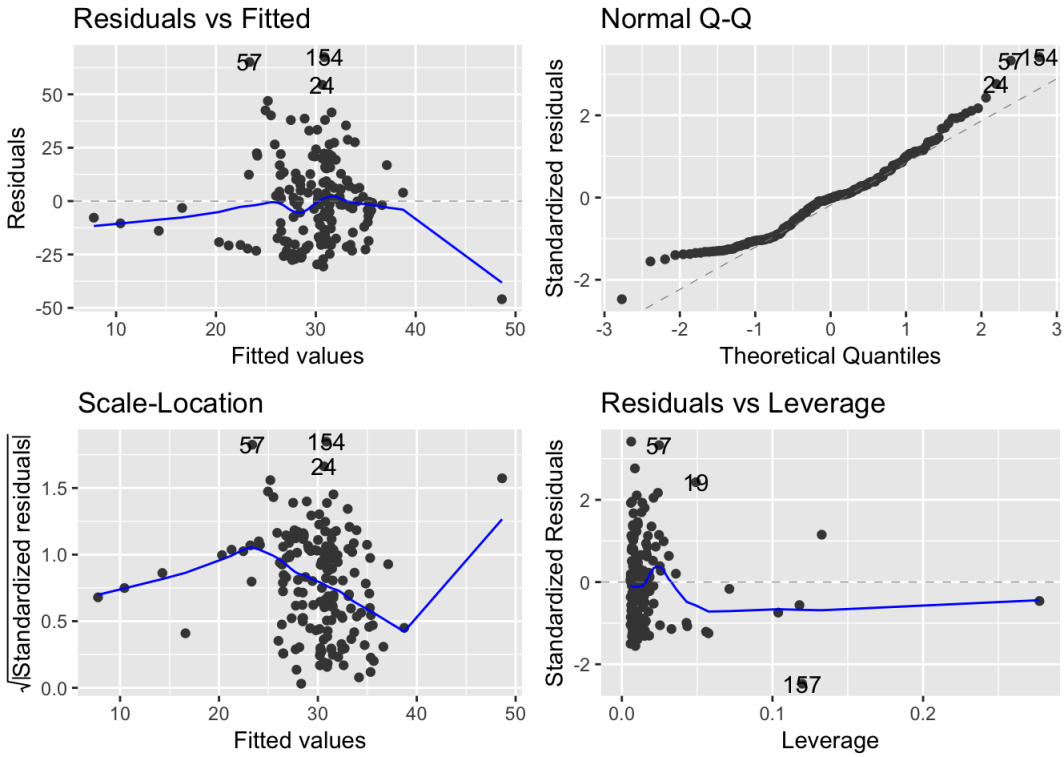
```
#diagnostic plots

#complex2 diagnostic plots

autoplot(complex2)
```




```
autoplot(sub2)
```



Residuals vs. Fitted (top left)

- In both complex2 and cub2, the Residuals vs. Fitted plots show a pattern that the values mostly are located in a specific area indicating possible non-linear relationships that the linear model has not captured. The vertical spread of the points is not the same throughout the whole graph, mostly in one area, looking somewhat like a triangle, indicating heteroscedastic errors.

Normal Q-Q (top right)

- In both complex2 and sub2, Normal Q-Q plot demonstrates that the residuals are not normally distributed because there are points that fall outside, and some above, of the dashed line.

Scale-Location (bottom left)

- Similarly to the Residuals vs. Fitted plots, both complex2 and sub2 Scale-Location plots indicate possible non-linear relationships that the linear model has not captured. Also shows heteroscedastic errors.

Residuals vs. Leverage (bottom right)

- In both complex2 and sub2, Residuals vs. Leverage plots there lies a point to the far right is a point of high leverage and contain some potential outliers indicating a bad fit.

ASSUMPTIONS:

- According to the textbook, OpenIntroStatistics, Multiple regression model generally depend on the following four assumptions: the residuals of the model are nearly normal, the variability of the residuals is nearly constant, the residuals are independent, and each variable is linearly related to the outcome.

CONCLUDE:

- Based on the F-test we rejected the null hypothesis that the simple model is better. Thus, chosing the complex2 model as the better choice. The expected negative correlational relationship between the outcome and the main predictor could be seen from the complex2 confidence interval -7.473069 and -2.744234, and this suggested the presence of statistical and practical significance. Unfortunately, however, neither model diagnostic plots suffice the multiple regression model assumptions. Therefore, the complex2 model nor the sub2 model are not good enough due to their shortcomings, indicating that the negatively correlational relationship might not be credible enough.

Conclusion

As expected, the models all showed a relationship with the deterioration of the environment and population growth, deforestation as the population grows and increasing pollution as the population grows. The coefficients for the primary predictor variable, population growth, told a consistent story, but this slight association is not causation. Besides, the models all contained shortcomings as shown in the diagnostic plots, therefore, should not be accepted as a model for prediction or indication. First, not all countries were included; accordingly, there was a bias, not accounting for the global picture. If the whole global picture is not shown, we may misinterpret the environmental impact these outcomes have. In other words, the data does not account for all the countries leading to a biased result, not considering the global population growth. Also, there are many factors related to population growth that vary from country to country. For example, the United States and Brazil share a similar population growth rate, the population growth in the United States is mainly due to immigration, but

the population growth in Brazil is due to domestic growth. In a word, “population growth rate” by individual countries do not tell us much information behind the number. As a result, I would conclude that to show a stronger association between population growth and environmental deterrence and come to any casual conclusion; one would need to collect unbias data and include more predictor variables in their model.