Simple Linear Regression Lab

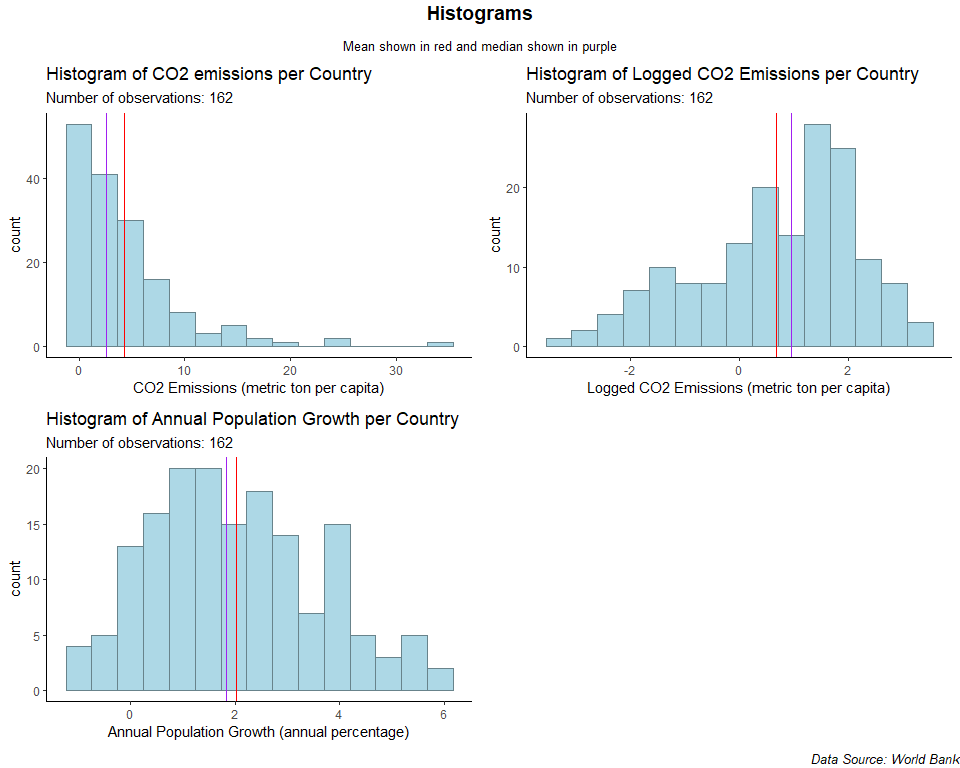
Francesca Chiappetta

9/29/2021

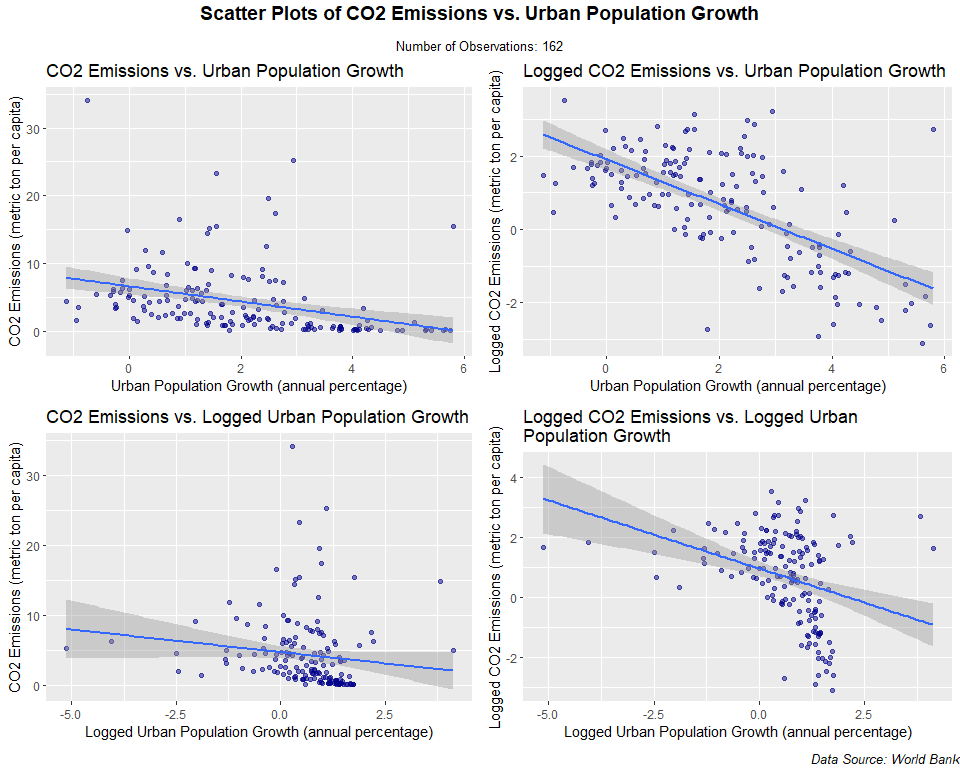
#Summary Statistics

Data of CO2 emissions in metric tons per capita and Urban Population Growth as annual percentage was taken from 162 countries was taken for. The range for CO2 emissions is 0.044485376 to 34.16324263 and the range for urban population growth is -1.123001709 to 5.79180675. The CO2 emission data have a strong positive skew of r summary\_co2\_urbangrowth\_df\_long$skew\_value[1], and a of r summary\_co2\_urbangrowth\_df\_long$mean\_value[1] that is larger than the median, r summary\_co2\_urbangrowth\_df\_long$median\_value[1] and a standard deviation of r summary\_co2\_urbangrowth\_df\_long$sd\_value[1]. The urban population growth data have a very small positive skew of r summary\_co2\_urbangrowth\_df\_long$skew\_value[2] a of r summary\_co2\_urbangrowth\_df\_long$mean\_value[2] that is larger than the median, r summary\_co2\_urbangrowth\_df\_long$median\_value[2] and a standard deviation of r summary\_co2\_urbangrowth\_df\_long$sd\_value[2]. The correlation coefficient is used to measure how strong the relationship is between CO2 emissions and urban population growth. The correlation coefficient of -0.34659615068 means that the two variables have a weak, negative linear relationship.

#Histograms



#Scatterplots

 Urban Population Growth is the **explanatory variable** and CO2 emissions is the **response variable** because I am interested to know whether an increase in urban population growth shows a positive linear relationship with CO2 emissions. There is a weak linear relationship between the two variables. log-linear graph displays the strongest linear relationship, with the data points more evenly scattered compared to the linear-linear, linear-log, and log-log scatter plots. The log-linear scatter plot shows a negative linear relationship.

#Linear Regression

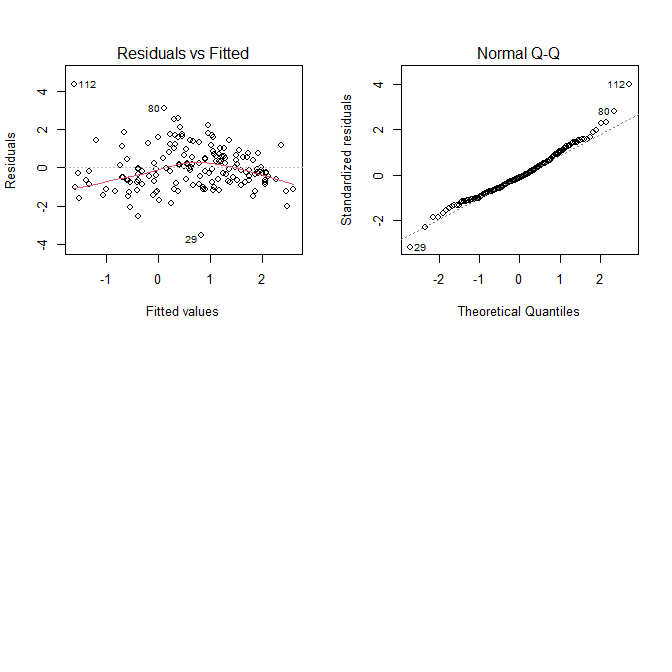
Only the dependent/response variable is log-transformed. Exponentiate the coefficient, subtract one from this number, and multiply by 100. This gives the percent increase (or decrease) in the response for every one-unit increase in the independent variable. Example: the coefficient is 0.198. (exp(0.198) – 1) \* 100 = 21.9. For every one-unit increase in the independent variable, our dependent variable increases by about 22%.

y = -45.55519925837 x + 1.90974995206

A one-unit increase in the explanatory variable (independent variable), Urban Population Growth (annual percentage), decreases the response variable (dependent variable), CO2 Emissions (metric tons per capita), by 45.55519925837%.

The r-squared is how well the regression model fits the observed data and generally, a higher r-squared indicates a better fit for the model. The r-squared value of 0.43072997029 reveals that 43.07299702901 % of the data fit the regression model. Although a “good” R square value depends on many factors, this R square value suggests that the linear model does not fit our data very well.

#Residual vs. Fitted Plots



## Run a residual vs fitted plot on the regression you ran above. Discuss the underlying assumptions of an OLS model and whether you think they are met in this context. (10 points).

**1.The model is linear in parameters.** The low R squared values suggests that this assumption is false, and the model does not fit the data.

**2. We have a random sample of size n (xi, yi), where i =1,2,3,4…n** This assumption requires that there is a random sampling of observations. FRANKIE CHECK ON THE DOCUMENTATION HERE

**3. The conditional mean should be zero, E(u|x) = 0** The curvature in the plot of residuals vs predicted Y shows us that the average error is not zero everywhere, suggesting non linearity and that the model does not fit the data.

*4. Sample outcomes of x are not all the same value.* Our sample outcomes of x are not all the same value, so this assumption holds true.

**5. The error u has the same variance given any value of the explanatory variable: var(u|x) = σ2** The plot of residuals against the fitted data show heteroskedasticity, meaning that the u does not have the same variances give any value of the explanatory variable. This assumption is violated.

**6. Errors are normally distributed** The second plot (normal Q-Q) is a normal probability plot and will give a straight line if the errors are distributed normally. The QQ plot of the residuals shows skewed residuals, meaning that our errors are not normally distributed and violate this assumption.