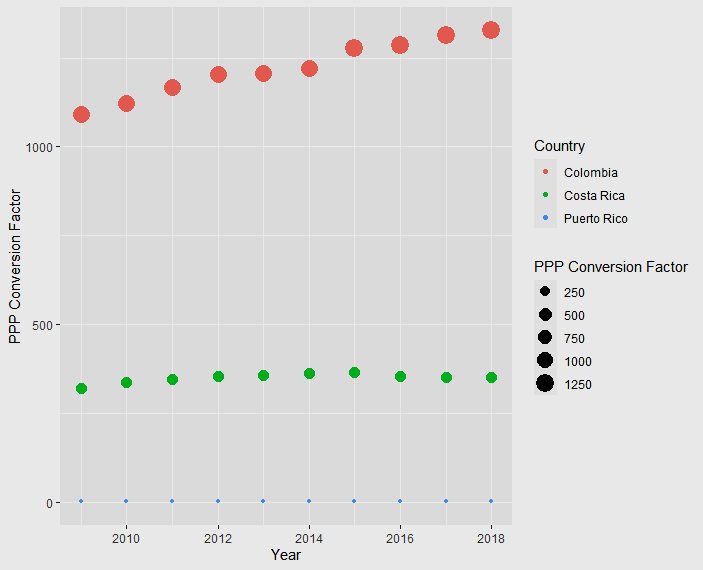
**Examining WBDI trends of Puerto Rico, Costa Rica, & Colombia**

**Business Question**

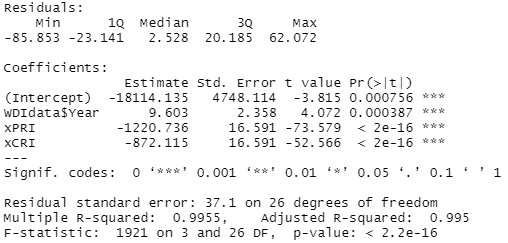
Is there a difference in the trend in a world bank development indicator between Puerto Rico, Costa Rica, and Colombia?

**GGPlot Exploring Business Question**

****

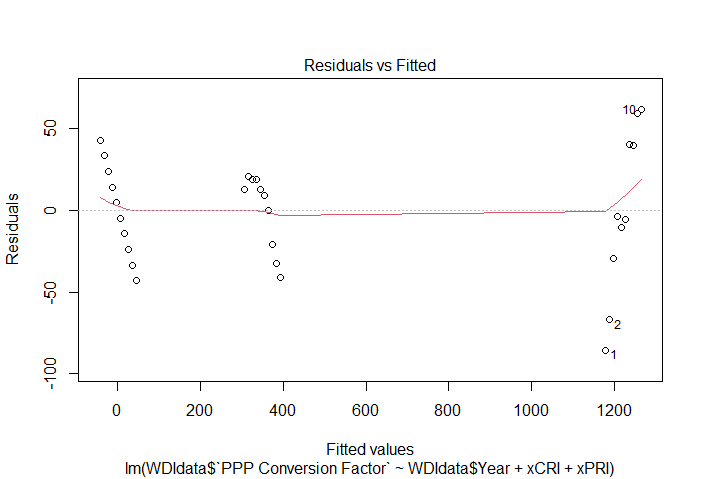
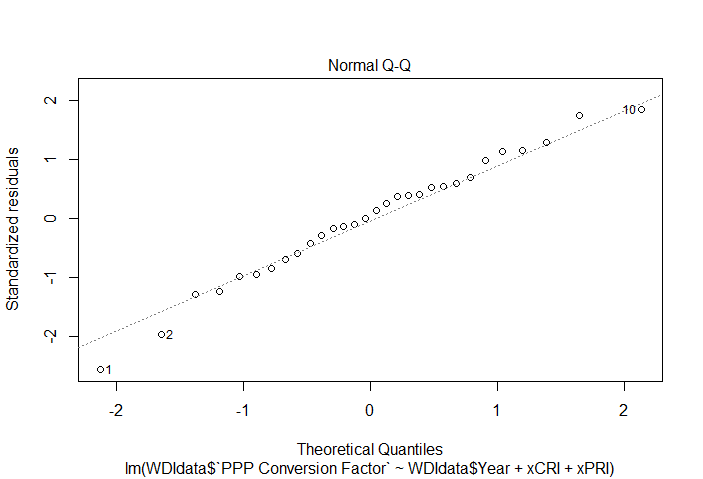
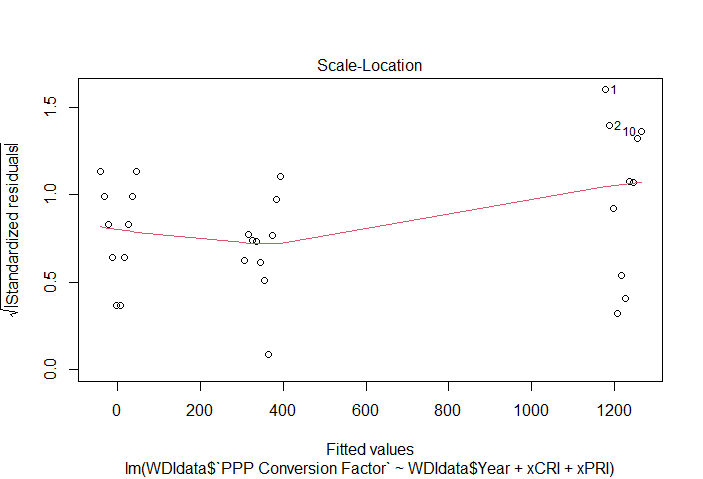
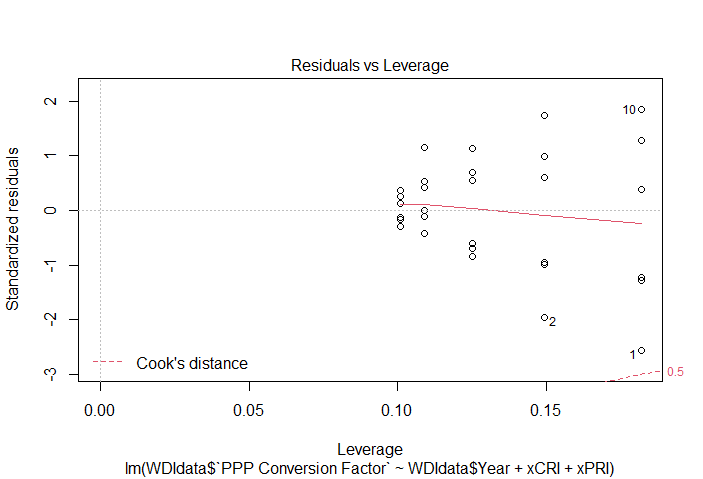
\*Based on the GGPlot above, there seems to be an obvious difference in the PPP Conversion Factor trends between Puerto Rico, Costa Rica, and Colombia. Over time, both Costa Rica and Puerto Rico have remained fairly stagnant with a slope of almost zero. On the other hand, Colombia’s PPP Conversion Factor seems to be increasing linearly between 2009 - 2018.

**Overall F Test**



* **Parameters**:
  + Y = PPP Conversion Factor
  + ß0 = Beta0 for Year
  + Indicator Variables: Puerto Rico, Costa Rica
    - ß1 = Slope for PRI (Puerto Rico)
    - ß2 = Slope for CRI (Costa Rica)
  + Base Variable: Colombia
* **Model**: yhat = -18114.135 + 9.603YEAR - 1220.736xPRI - 872.115xCRI
* **H0**: ß1 = ß2 = ß3
* **HA**: At least one ßi =/= zero
* **Assumptions:**

|  |  |
| --- | --- |
| **Assumption** | **Condition** |
| 1. Trend is linear. | The scatterplots show a linear enough pattern between the explanatory variables and response variables on the normal quantile plot. |
| 1. Errors are independent. | In the residual vs. fitted plot, there is no pattern. |
| 1. Variability of errors is constant. | In the residual vs. predicted plot, there is no funnel shape present. |
| 1. Errors follow a normal model. | In the normal quantile plot, the errors are close to normal with a few error outliers.  ***Limitation: Based on the normal quantile plot, the regression data set has three outliers at points 1, 2, and 10.*** |

* + Residual vs. Predicted Plot:
    - 
  + Normal Quantile Plot
    - 
  + Fitted Values vs. √Standardized Residuals
    - 
  + Residuals vs. Leverage
    - 
* **Test Statistic**: 1921 → ***High test statistic - helpful in predicting y!***
* **R2**: 0.995 → ***Higher R2 value - helpful in predicting y!***
* **P-value**: P-value is almost zero - ***helpful in predicting y!***
* **Conclusion**: Reject the H0, there is a difference in the trends in the PPP Conversion Factor between the countries. We can safely conclude at least one of these variables is significant and at least one of the countries is different. Year is significant even when ß1 and ß2 are in the model.

**LM Function with a Squared Term**

* New Adjusted R2 = -0.07089
* After incorporating the squared term into the LM function, the Adjusted R2 value decreased substantially from 0.995. With a lower R2, the original multiple linear regression model was a more accurate model.

**Next Steps**

* In future modeling of these relationships, since the Adjusted R2 value decreased when adding a squared term, there is no evidence suggesting we should attempt to model these relationships by adding a cubic term because the adjusted R2 would further decrease; therefore, leaving us with an even less accurate model. Again, due to the accuracy of the multiple linear regression model (as a result of the high Adjusted R2 value), we will consider that to be the most accurate model.
* To increase the accuracy of our original model, I would further remove Puerto Rico from the model.

**RStudio Code**

**##Invoking the TidyVerse package in order to use functions from that package for the assignment.**

* library(tidyverse)

**##Creating a GGPlot of Y vs. Time(x) with each different country denoted as a different color and different size.**

* ggplot(data=WDIdata)+geom\_point(mapping=aes(x=Year, y=`PPP Conversion Factor`, color=Country, size=`PPP Conversion Factor`))

**##Creates an indicator variable based on Code. x\_CR when the country is Costa Rica, and zero otherwise.**

* xCRI<-ifelse(WDIdata$Code == "CRI", 1, 0)

**##Creates an indicator variable based on Code. x\_PR when the country is Puerto Rico, and zero otherwise.**

* xPRI<-ifelse(WDIdata$Code == "PRI", 1, 0)

**##Creates a multiple regression equation below to predict PPP Conversion Factor based on Country. The summary yields the overall F test and also yields the regression equation.**

* Multiple <-lm(WDIdata$`PPP Conversion Factor`~WDIdata$Year+xCRI+xPRI)
* summary(Multiple)

**##Creates regression plots which aid in determining regression model accuracy.**

* plot(Multiple)

**##Adds a squared term into the multiple linear regression model and displays the summary and plots for the new model.**

* SquaredLM<-lm(`PPPConversionFactor`~poly(Year,degree=2,raw=TRUE),data=WDIdata)
* summary(SquaredLM)
* plot(SquaredLM)