Data

The data utilized was a pharmaceutical spending dataset grabbed from the Organization of Economic Cooperation and Development database. The OECD is a good place to find various datasets pertaining to its member countries spending. The dataset contains information on 496 observations six categories for 34 OECD member countries: time, location, pharmaceutical expenditure, healthcare expenditure (both measured in US dollars per capita), nurses per 1000 people, and doctors per 1000 people. Each of these values is measured in R as an integer, with the exception of location, which had to be changed into a factor. The timespan of this data can go as far back as 1970, but for the sake of having the most information in our time range, we will be working with the timespan of 1980 until 2016. The dataset also includes "flag codes" for some countries as well, but that was cleaned out for the sake of easily sorting and coding the data in our analysis software, Rstudio, the flag codes were not a factor that needed to be observed due to too many "NA" values.

The OECD dataset allows for the amount of graduates to be added to the dataset as well, but we will be working with those first four variables. The relevance of our dataset is a tracking of pharmaceutical and healthcare prices, along with the locations and professionals in the medical field to see if there is a ground-level impact that medical professionals have on rising pharmaceutical prices. This is tied to our literature in a way that allows for a more comprehensive look at what would affect pharmaceutical prices in a way that leaves no stone unturned, so to speak. Leading into the descriptive statistics of the data set, 3 figures are included: figure 1 has the descriptive statistics of the natural dataset with no omitted variables, figure 2 are the descriptive statistics with some omitted variables that did not fit our first model,

and figure 3 is for the model that works with the top countries due to their prices being so high that the 29 other countries cause for higher standard error.

With regards to figure 1: The measures of skewness are mainly towards the right end on different levels (due to being positive, further positive is further skewed to the right) while the kurtosis varies around the normal expected value which is usually 3. The distributions for health expenditure and pharmacy expenditure are greater, meaning they are leptokurtic, which produce more outliers than a normal distribution and makes sense with the amount of countries that are collected by the OECD. The kurtosis measurements for the amount of doctors is close to the normal distribution's level of 2.96, but both that and the measure for nurses are platykurtic, which tends to produce less outliers than the normal distribution. With regards to figures 2 the mean values for health and pharmaceutical expenditures are lower. Figure 3 is interesting due to having mostly negative kurtosis, which means that the distribution of data has lighter tails and a flatter peak. You can see in each figure our response variable is not entirely in a normal range, but based on the law of large numbers and central limit theorem, the distributions will be normal around the true parameter values. With the regression for the largest dataset remains nearly normal, and valid under OLS and central limit theorem so long as it is within true parameters of the model estimation.

Methodology

As previously mentioned, the statistical software used is R(studio) with the goal of creating a linear regression model that ties our independent variables (healthcare expenditure, doctors/1000, and nurses/1000) to our dependent variable. With regression and data manipulation in mind, the packages in use are: caret (classification and regression training),

e1017(for descriptive statistics), ggplot2 (visuals), gvlma (to evaluate using linear assumptions), magrttr(helps with ggplot), plyr (for data splitting), and the base packages for R. Our training and testing data is split on a 75%/25% ratio for our model for 3 datasets. Healthset is our natural set, healthset 2 has only our top 5 countries, and healthset 3 removes some of the outlier countries from the natural set. Along with the descriptive statistics, figure 4 gives a look at correlation values between variables with exception to location. The closer to 1, the stronger the correlation.

In order to see which of our models are viable for linear regression, we have to test each model under the OLS assumptions. The gvlma package gives a value and p value reading for the global stat, skewness, kurtosis, link function, and heteroscedasticity for each model that is put through it. When each model was run through the tests, the only model that OLS held true for was the second. Therefore our top 5 countries will be used to build a model, figure 5 shows a time series of their pharmaceutical expenditures. The other countries that were not part of the top 5 set were not depicted due to how large the plot would be. Below is the formula for the model as shown here and the read from R in figure 6:

PharmaExp =-8960+ 4.375(TIME) -595.6(CHE) -117.9(DEU) -154.9(JPN) -232.7(USA)

(11200) (5.657) (118.5) (82.34) (31.44) (73.11)

+0.07251(HealthExp) -273.3(Doctors/1000p)+ 133.7(Nurses/1000p)

(0.0952) (66.79) (16.22)

Some issues that have been run into with the model is Canada (CAN) has been naturally omitted via the software, we will see in the results if this has an effect on the test prediction, an issue with omitting Germany (DEU), and time not being measured as a significant factor even though it is on a time series. Outside of that, the prediction accuracy of the data went pretty well.

The correlation of the prediction in the model will be shown in the next section.

Figures and Results

PharmaExp =-8960+ 4.375(TIME) -595.6(CHE) -117.9(DEU) -154.9(JPN) -232.7(USA)

(11200) (5.657) (118.5) (82.34) (31.44) (73.11)

+0.07251(HealthExp) -273.3(Doctors/1000p)+ 133.7(Nurses/1000p)

(0.0952) (66.79) (16.22)

Above is the regression. The Adjusted R-Squared value is at 0.9737 , which is the highest of all possible models, the F-Statistic is 200.4 on 8 and 35 degrees of free(figure 6). The AIC (438.1) and BIC(455.9), which are relatively low compared to the other models that were used. The mean absolute percentage error is 5.37%, and the mean square error is calculated as 784.87. Figure 7 shows the comparison of the prediction in comparison to the actual testing set (both of 15 points), with a correlation value of 0.949, the models is a success in being able to estimate pharmaceutical prices within the top 5 highest spending countries.

1.

```
     TIME         LOCATION     HealthExp        PharmaExp
Min.   :1980   ISL     : 36   Min.   : 238.2   Min.   :  56.97
1st Qu.:2001   AUS     : 33   1st Qu.:1341.1   1st Qu.: 267.43
Median :2006   CZE     : 24   Median :2372.7   Median : 393.86
Mean   :2005   DNK     : 23   Mean   :2683.2   Mean   : 408.53
3rd Qu.:2011   ESP     : 23   3rd Qu.:3722.3   3rd Qu.: 524.35
Max.   :2016   HUN     : 22   Max.   :9035.5   Max.   :1081.40
               (Other):335
DoctorsPer1000    NurPer1000
Min.   :1.090   Min.   : 0.930
1st Qu.:2.440   1st Qu.: 6.100
Median :3.030   Median : 8.480
Mean   :2.997   Mean   : 8.765
3rd Qu.:3.442   3rd Qu.:11.140
Max.   :5.100   Max.   :17.950
```

```
> skewness(healthset$HealthExp)
[1] 0.9262065
> skewness(healthset$PharmaExp)
[1] 0.6183313
> skewness(healthset$DoctorsPer1000)
[1] 0.008988921
> skewness(healthset$NurPer1000)
[1] 0.2170072
```

```
> kurtosis(healthset$HealthExp)
[1] 3.635582
> kurtosis(healthset$PharmaExp)
[1] 3.489944
> kurtosis(healthset$DoctorsPer1000)
[1] 2.952249
> kurtosis(healthset$NurPer1000)
[1] 2.52415
```

Descriptive statistics for the first dataset, with everything involved. Although the measurements were relatively close to normal, the model based on this data failed OLS.

2.

```
      TIME          LOCATION     HealthExp      PharmaExp      DoctorsPer1000    NurPer1000
 Min.   :2000   DEU    :16   Min.   :2105   Min.   : 381.5   Min.   :1.980   Min.   : 8.43
 1st Qu.:2005   USA    :15   1st Qu.:3511   1st Qu.: 559.8   1st Qu.:2.350   1st Qu.: 9.68
 Median :2009   CAN    :13   Median :4502   Median : 694.3   Median :2.510   Median :10.74
 Mean   :2008   CHE    : 8   Mean   :4902   Mean   : 706.8   Mean   :2.908   Mean   :11.31
 3rd Qu.:2012   JPN    : 7   3rd Qu.:6072   3rd Qu.: 815.9   3rd Qu.:3.580   3rd Qu.:11.89
 Max.   :2015   AUS    : 0   Max.   :9036   Max.   :1081.4   Max.   :4.200   Max.   :17.95
```

```
> skewness(healthset2$HealthExp)
[1] 0.5884359
> skewness(healthset2$PharmaExp)
[1] 0.1316597
> skewness(healthset2$DoctorsPer1000)
[1] 0.4736636
> skewness(healthset2$NurPer1000)
[1] 1.314638
```

```
> kurtosis(healthset2$HealthExp)
[1] -0.659025
> kurtosis(healthset2$PharmaExp)
[1] -0.9805638
> kurtosis(healthset2$DoctorsPer1000)
[1] -1.432472
> kurtosis(healthset2$NurPer1000)
[1] 0.9081813
```

Dataset 3 was based on a cleaned version of dataset 1 that removed the locations that was lower than 0.5 alpha in significance. This set failed OLS as well and was not fit for linear regression.

3.

```
> summary(healthset3)
      TIME          LOCATION     HealthExp      PharmaExp      DoctorsPer1000    NurPer1000
 Min.   :1980   ISL    : 36   Min.   : 238.2   Min.   :  56.97   Min.   :1.090   Min.   : 0.930
 1st Qu.:2001   AUS    : 33   1st Qu.:1332.6   1st Qu.: 257.36   1st Qu.:2.380   1st Qu.: 6.130
 Median :2006   CZE    : 24   Median :2379.9   Median : 375.82   Median :2.960   Median : 8.890
 Mean   :2005   DNK    : 23   Mean   :2736.7   Mean   : 397.59   Mean   :2.974   Mean   : 8.901
 3rd Qu.:2011   ESP    : 23   3rd Qu.:3773.4   3rd Qu.: 511.59   3rd Qu.:3.560   3rd Qu.:11.360
 Max.   :2016   KOR    : 20   Max.   :9035.5   Max.   :1081.40   Max.   :5.100   Max.   :17.950
                (Other):256
```

```
> skewness(healthset3$HealthExp)
[1] 0.9271759
> skewness(healthset3$PharmaExp)
[1] 0.7954077
> skewness(healthset3$DoctorsPer1000)
[1] 0.09176547
> skewness(healthset3$NurPer1000)
[1] 0.1021836
```

```
> kurtosis(healthset3$HealthExp)
[1] 0.4803113
> kurtosis(healthset3$PharmaExp)
[1] 0.6994457
> kurtosis(healthset3$NurPer1000)
[1] -0.5545768
> kurtosis(healthset3$DoctorsPer1000)
[1] -0.4298985
```

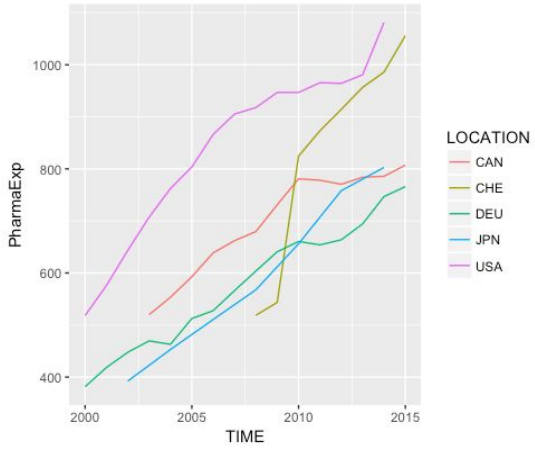
Although the data from our top 5 countries seems odd. When ran through R using the gvla package, this is the only dataset that fills OLS assumptions for linear regression.

4.

```
> cor(healthset2)
                    TIME HealthExp PharmaExp DoctorsPer1000 NurPer1000
TIME          1.00000000 0.5072429 0.6689116      0.3918010 0.06962438
HealthExp     0.50724289 1.0000000 0.8189234      0.3343370 0.63277351
PharmaExp     0.66891158 0.8189234 1.0000000      0.2943713 0.34073287
DoctorsPer1000 0.39180101 0.3343370 0.2943713      1.0000000 0.34760496
NurPer1000    0.06962438 0.6327735 0.3407329      0.3476050 1.00000000
```

Shown are the correlation numbers for our best dataset, it should be noted that there is high correlation between health expenditure and pharmaceutical expenditures.

5



A graph of the top 5 countries expenditures. These were chosen due to their similarity in being vastly more expensive than the 29 other countries.

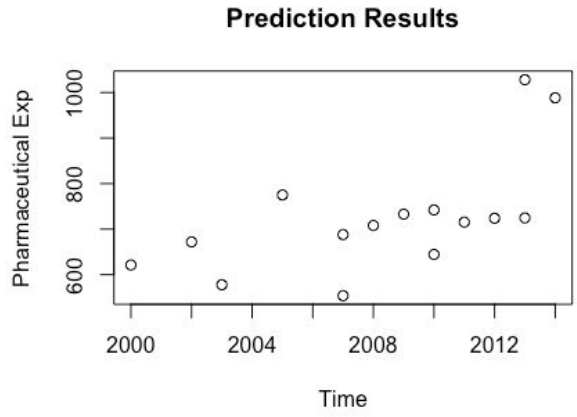
6.

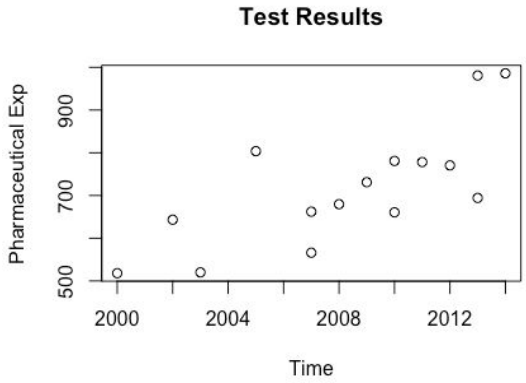
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.960e+03  1.120e+04  -0.800 0.429004
TIME            4.375e+00  5.657e+00   0.773 0.444452
LOCATIONCHE    -5.956e+02  1.185e+02  -5.026 1.48e-05 ***
LOCATIONDEU    -1.179e+02  8.234e+01  -1.432 0.161058
LOCATIONJPN    -1.549e+02  3.144e+01  -4.926 2.01e-05 ***
LOCATIONUSA    -2.327e+02  7.311e+01  -3.183 0.003057 **
HealthExp       7.251e-02  1.952e-02   3.714 0.000707 ***
DoctorsPer1000 -2.733e+02  6.679e+01  -4.092 0.000239 ***
NurPer1000      1.337e+02  1.622e+01   8.238 1.04e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.4 on 35 degrees of freedom
Multiple R-squared:  0.9786,    Adjusted R-squared:  0.9737
F-statistic: 200.4 on 8 and 35 DF,  p-value: < 2.2e-16
```

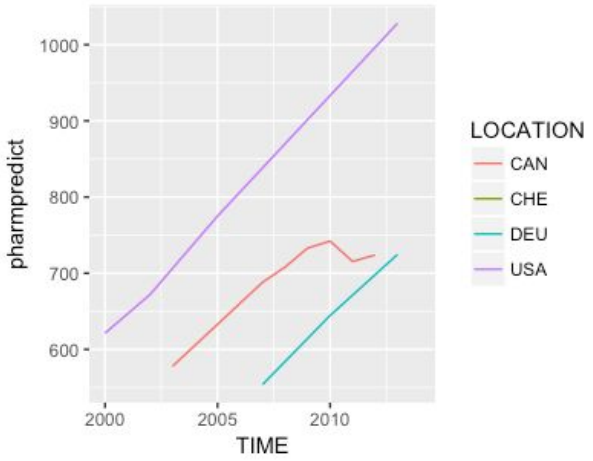
Displayed is the summary reading for the m2 model. The intercepts for and Germany and time

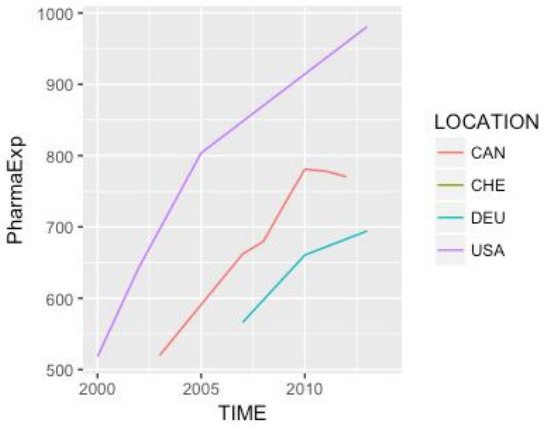
should be included even though the reading says it is not significant if only to avoid a possible domino effect on other variables. We are working with a smaller observed group of data in this top 5 set, so removing any variable can lead to skewed results. It is possibly that the later years (figure 5) of Germany can skew with regards, but out of each model the adjusted R-Squared is as high as it's been with the lowest standard error.





7.





Both the scatterplot and the time series graph represent the same information, the final graphs were only able to have predictions, shown on the left, for 3 of the five due to the data splitting. That being done though there is high correlation to the actual test, shown on the right.