**Stage 4 Member Task Report-Kyle Killworth**

**Part 1: Utilize Linear and Non-Linear (polynomial) regression models to compare trends for Indiana**

COVID Cases

**A picture containing text, screenshot

Description automatically generated**

COVID Deaths

A picture containing text, screenshot

Description automatically generated

As we can see, the nonlinear models for cases appear to be nearly identical to the linear model based on both the graph and RMSE, so any further analysis likely should stick to the linear model. Interestingly, this differs greatly from graphs you'll see below for the 5 counties we pulled out. I believe it may have to do with some counties underreporting or not reporting cases at all, leading to far lower values than intended, pulling the regression lines down.

It's even more pronounced with the deaths, so my previously mentioned theory above applies here as well. Again, the linear model seems to be the best one to use based on the graph and RMSE.

**Part 2: Utilize Linear and Non-Linear (polynomial) regression models to compare trends for top 5 counties for COVID cases in Indiana**

|  |  |
| --- | --- |
| **County\_Name** | **Covid Cases** |
| st. joseph county | 69301 |
| hamilton county | 79781 |
| allen county | 101761 |
| lake county | 104518 |
| marion county | 220797 |

The counties above are the top 5 in terms of overall COVID cases for Indiana, so they were the ones I used for this part of the project.

Allen County COVID Cases

Chart, line chart

Description automatically generated

Allen County COVID Deaths

Chart, scatter chart

Description automatically generated

Now we can see the impact of variance on the nonlinear models for cases. An argument could be made to use the third order nonlinear model of the linear model, depending on your definition of 'significant decrease in RMSE'. For deaths, it appears only the third order nonlinear model was able to truly become nonlinear, as we can only barely see the linear model next to the second-order nonlinear model. However, this is still an improvement over the Indiana deaths graph.

Hamilton County COVID Cases

Chart

Description automatically generated

Hamilton County COVID Deaths

Chart, scatter chart

Description automatically generated

Hamilton is similar to Allen when it comes to cases. Again, an argument could be had for the third-order nonlinear model. For deaths, it seems like the non-linear models aren't much different, similar to Indiana as a whole. RMSE is nearly identical for all models.

Lake County COVID Cases

Chart

Description automatically generated

Lake County COVID Deaths

Chart

Description automatically generated with medium confidence

For cases, a similar graph to the other counties, and an argument could be made for the third-order non-linear model, though all trends look to be a tad flatter. For deaths, we can actually see the linear model clearly without being overshadowed by the second-order nonlinear model, though the differences between the three are still very slight.

Marion County COVID Cases

Chart

Description automatically generated

Marion County COVID Deaths

A picture containing text, screenshot

Description automatically generated

Same trends for cases that we've seen from the other counties, though not as flat as Lake. Has the biggest decrease in RMSE from linear to third order nonlinear. We're back to not really being able to distinguish between the linear and second-order nonlinear models for deaths, and linear continues to be the best pick.

St. Joseph County COVID Cases

Chart

Description automatically generated

St. Joseph County COVID Cases

A picture containing text, screenshot

Description automatically generated

Not surprisingly, cases look the same as they did for the other four counties, and even the RMSE is similar to Marion for the three models. Interestingly for deaths, the second-order nonlinear model here is instead being overshadowed by third-order, instead of overshadowing the linear model. Other than that, the same trends as we've been seeing in the other counties.

**Part 3: Utilize the hospital data to calculate the point of no return for a state. Use percentage occupancy / utilization to see which states are close and what their trend looks like**

The next stage of the project is to look at hospital bed occupancy, and compare that to COVID numbers, to see if we hit a point of no return. COVID cases are a point indicator for this, as the majority of people who get COVID do not need to go to the hospital. Instead, we'll use COVID deaths, though one could argue this is also not a great predictor. We will be comparing the deaths in our linear model to a threshold based on the number of ICU beds available in a state. We looked at Indiana, Illinois, Kentucky, Michigan, and Wisconsin.

Indiana ICU Beds Point of No Return

Shape

Description automatically generated

As we can see, even when just looking at ICU beds, the regression line for COVID deaths for Indiana gets nowhere near the threshold, even when increasing the prediction for a whole year. So, instead of graphing it, let's try solving for it using the regression coefficients.

Indiana COVID Death Linear Regression Coefficients

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 0.1610 | 0.012 | 13.533 | 0.000 | 0.138 | 0.184 |
| **Days\_Infection** | 0.0004 | 2.75e-05 | 15.325 | 0.000 | 0.000 | 0.000 |

Now that we have the coefficients, we can calculate the number of days from infection we predict will hit the point of no return using the regression model equation, but solving for x instead of y. After solving for x, and dividing by 365 to scale it to years, we get approximately 9,581 years. So apparently it will take approximately 9581 years to hit the point of no return. The other four states have similar graphs and results, so I won’t provide those graphs. Here is a table with the linear regression calculations.

|  |  |
| --- | --- |
| **State** | **Point of No Return** |
| Indiana | 9,581 |
| Illinois | Not possible due to Infection coefficient being negative |
| Michigan | 12,622 |
| Kentucky | 7,567 |
| Wisconsin | 6,396 |

**Part 4: Perform hypothesis tests on questions identified in Stage II**

The last step of the member task is to conduct some hypothesis tests. As a reminder, these were the three hypotheses:

* Hypothesis 1: Does an increase in foreign born residents lead to higher counts of COVID cases or deaths?
* Hypothesis 2: Does having a higher number of residents with no diploma lead to higher counts of COVID cases or deaths?
* Hypothesis 2: Does having a higher number of residents with a Bachelor's degree lead to higher counts of COVID cases or deaths?

For the first hypothesis, we used a two-tailed two-sample t-test, with significance level at α=.05, followed by one-tailed two-sample t-tests. The null hypothesis is that the number of foreign born residents has no bearing of covid cases counts. Based on the result of the two-tailed two-sample t-test, we can reject the null hypothesis, as our p-value is less than significance level at α=.05, with a value of 0.034. We now need to check if this is still true with a one-tailed two-sample t-test. With one-tailed tests, our significance level needs to be halved to .025. We can still reject the null hypothesis with the one-tailed test looking at the positive end of the distribution, with a p-value of 0.017. An increase in foreign born residents leads to a significant increase in COVID cases. Like with cases, we can also reject the null hypothesis for deaths based on the p values of the two tailed test (0.029) and the positive one-tailed test (0.015). An increase in foreign born residents leads to a significant increase in COVID deaths.

For our second hypothesis, we used a two-tailed two-sample t-test, with significance level at α=.05, followed by one-tailed two-sample t-tests. The null hypothesis is that the number of residents with no diploma has no bearing of covid cases counts. As our p-value is greater than 0.05 at 0.06, we fail to reject the null hypothesis for cases. Having higher number of residents with no diploma has no impact on COVID cases. Unlike with the cases, deaths are significant both at the two-tailed test (p-value 0.044) and the positive one-tailed test (p-value 0.022). We can reject the null hypothesis. Having higher number of residents with no diploma leads to an increase in COVID deaths.

For our third hypothesis, we used a two-tailed two-sample t-test, with significance level at α=.05, followed by one-tailed two-sample t-tests. The null hypothesis is that the number of residents with a Bachelor's degree has no bearing of covid cases counts. With a p-value of 0.04 for the two-tailed test, and a p-value of 0.02 for the positive one-tailed test, we can reject the null hypothesis for cases. Having a higher number of residents with a Bachelor's degree leads to a significant increase in COVID cases. Like with cases, we can reject the null hypothesis for deaths with p-values of 0.038 for the two-tailed test and 0.019 for the one-tailed test. Having a higher number of residents with a Bachelor's degree leads to a significant increase in COVID deaths.