**4.1: Transforming to Achieve Linearity**

Sometimes (actually, pretty often) we will encounter data that has a non-linear relationship. In order to analyze this non-linear data using our linear tools, we have to transform the data with a mathematical function for the relationship between the two variables to become approximately linear.

Ex: There has been a zombie outbreak in the USA. Mrs. Skaff was able to gather the following data before turning into Zombie herself ☹ You (as a class of statisticians) have been asked to predict the # of Zombies that will be present 28 days after the initial outbreak.

|  |  |
| --- | --- |
| # days | # zombies |
| 0 | 1 |
| 1 | 2 |
| 2 | 3 |
| 3 | 9 |
| 4 | 17 |
| 5 | 40 |
| 6 | 75 |
| 7 | 130 |
| 8 | 200 |
| 9 | 450 |
| 10 | 880 |
| 11 | 1600 |

It looks exponential, we can verify this by calculating the ratio between the values at each step. If is the same for each pair of y values, we are looking at **\_\_\_\_\_\_\_\_\_\_\_** growth or decay.

It’s obviously not linear, but if you look at the residual plot and r2 we can gather more evidence to support our claim. (Sometimes it is less obvious in the scatterplot)

Exponential functions have the form:

To make this linear, we will can take the logarithm (or natural log) of each side.

**Notice that we took the natural logarithm of the y’s, but not the x’s.** We can transform our data using this knowledge.

|  |  |  |
| --- | --- | --- |
| # days | # z’s | ln(#zombies) |
| 0 | 1 | 0 |
| 1 | 2 | 0.693147 |
| 2 | 3 | 1.098612 |
| 3 | 9 | 2.197225 |
| 4 | 17 | 2.833213 |
| 5 | 40 | 3.688879 |
| 6 | 75 | 4.317488 |
| 7 | 130 | 4.867534 |
| 8 | 200 | 5.298317 |
| 9 | 450 | 6.109248 |
| 10 | 880 | 6.779922 |
| 11 | 1600 | 7.377759 |

We not have a linear equation that represents this data!!!

ln (number\_zombies) = 0.055 + 0.676(number\_days)

Let’s answer our question: How many zombies should we expect 28 days later?

Alternatively, we could “untransform” our linear equation to find the appropriate exponential equation.

ln (number\_zombies) = 0.055 + 0.676(number\_days)

We will look at one other non-linear relationship that often occurs.

On July 31st 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. They named the possible planet Xena. Xena is at an average distance of 9.5 billion miles. Assuming Xena is a planet, what would you predict its period of revolution to be? (FYI…the discovery of Xena led to Pluto’s demotion ☹ and it was decided that both Xena (now named Eris) and Pluto were “dwarf planets”)

|  |  |  |
| --- | --- | --- |
| Planet | Distance from Sun in millions of miles | Period of revolution (earth years) |
| Mercury | 36 | 0.241 |
| Venus | 67 | 0.615 |
| Earth | 93 | 1.000 |
| Mars | 142 | 1.881 |
| Jupiter | 484 | 11.862 |
| Saturn | 887 | 29.456 |
| Uranus | 1784 | 84.070 |
| Neptune | 2796 | 164.810 |
| Pluto | 3666 | 248.530 |

It’s definitely not linear… It’s also definitely not exponential…

Power law model:

Notice that our linear model includes taking the natural logarithm of **BOTH** x and y. Let’s apply this and transform our planet data….

|  |  |  |  |
| --- | --- | --- | --- |
| Distance from Sun (au) | Ln(distance) | Period of revolution (earth years) | Ln (years) |
| 36 | 3.583519 | 0.241 | -1.42296 |
| 67 | 4.204693 | 0.615 | -0.48613 |
| 93 | 4.532599 | 1.000 | 0 |
| 142 | 4.955827 | 1.881 | 0.631804 |
| 484 | 6.182085 | 11.862 | 2.47334 |
| 887 | 6.787845 | 29.456 | 3.382898 |
| 1784 | 7.486613 | 84.070 | 4.43165 |
| 2796 | 7.935945 | 164.810 | 5.104793 |
| 3666 | 8.206856 | 248.530 | 5.515564 |

Xena is at an average distance of 9.5 billion miles (9500 million). Assuming Xena is a planet, what would you predict its period of revolution to be?