Luke Watson

11-11-24

**Exercise 9**

*Part One*

In this exercise, I explored non-linear curve fitting in R. To start, I was instructed to fit a linear and polynomial regression to a subset of the Anscombe’s quartet dataset. Using a simple linear regression, I fit the Anscombe data with x1 as a predictor and y2 as the outcome. I repeated the same syntax with a spline fit. Then, I fit y2 with a polynomial regression, using x1 + x12 as predictors. The results and model fits are compared in Table 1. The polynomial regression provided the best fit to the data according to AIC (*AICquadratic* = -105, *AIClinear* = 40). Since these models were nested, I could also conduct an F-test to compare the model fits statistically. The quadratic regression fit the data better according to this test (*F*(1, 8) = 4925016, *p* < .001), matching the judgment of the AIC.

A visual inspection of the fits over the raw data shows the relationship is clearly non-linear and is likely why the polynomial quadratic regression fits the data much better (Figure 1). The quadratic regression even outperforms the spline.

**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model/Term | Estimate | SE | t | p | AIC |
| Spline |  |  |  |  | -103 |
| Intercept | 3.10 | .002 | 1951 | <.001 |  |
| X1(1) | 5.89 | .005 | 1221 | <.001 |  |
| X1(2) | 7.56 | .004 | 2119 | <.001 |  |
| X1(3) | 5.00 | .002 | 2115 | <.001 |  |
| Linear |  |  |  |  | 40 |
| Intercept | 3.00 | 1.130 | 2.67 | .026 |  |
| X1 | .50 | .120 | 4.24 | .002 |  |
| Quadratic |  |  |  |  | -105 |
| Intercept | -6.00 | .004 | -1385 | <.001 |  |
| X1 | 2.78 | .001 | 2674 | <.001 |  |
| X12 | -.13 | .000 | -2219 | <.001 |  |

*Note.* Comparison of the model fit and individual parameter values retrieved from the spline, linear, and quadratic polynomial regression. The spline is an exploratory model used as a baseline to compare the subsequent linear and quadratic model fits to.

**Figure 1**

A

B

A red line with numbers and a curve

Description automatically generatedA graph of a line

Description automatically generated with medium confidence

C

A red line graph with numbers and points

Description automatically generated

*Figure 1* – Visual comparison of model fits for linear (A), quadratic polynomial (B), and spline (C). The underlying data are plotted as points along X1 and show a clear non-linear relationship. Given how well the quadratic model fit, this relationship is likely quadratic in nature. Error ribbons represent 95% C.I. (On panel B, the error was so small that the ribbons cannot be seen).

*Part Two*

After fitting the Anscombe data with a linear and polynomial regression, I repeated the quadratic model as a non-linear regression using the nls command in R. Each parameter (X1, X12, and the intercept) has a coefficient (A, B, C) to fit the quadratic equation below to the data (Equation 1). I used a starting value of 1 for each coefficient (A, B, C) and let the model iterate to find the best fitting values for them. The results are summarized in Table 2. I encourage you to compare the values to the model parameters of the quadratic regression above, you will find that they are the same. A visual representation of the relationship is shown in Figure 2, although, this again is no different than the polynomial fit displayed in panel B of Figure 1.

**Equation 1**

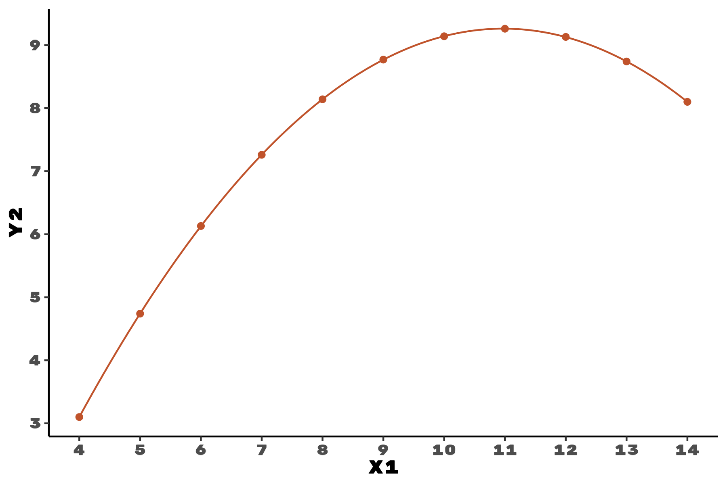
*Equation 1* – a can influence the shallowness or narrowness of the opening of the curve with lower values indicating wider curves and negative values indicating downward turning curves. b can move the curve around the origin and shift where the curve intercepts the x-axis. c is where the curve intercepts the y-axis.

**Table 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | SE | t | p |
| C (intercept) | -6.00 | .004 | -1385 | <.001 |
| B | 2.78 | .001 | 2674 | <.001 |
| A | -.13 | .000 | -2219 | <.001 |

*Note*. Coefficient values for the non-linear quadratic analysis of the Anscombe data. The coefficients were rearranged to display in the same format as the quadratic polynomial regression in Table 1 and show off how they are the same.

**Figure 2**



*Figure 2* – Fitted curve from the nonlinear quadratic regression over the raw data points. Looks the same as the curve in panel B of Figure 1.

*Part Three*

After getting some nonlinear analysis under my belt, I moved on to fitting nonlinear models to actual data. The data I used was a repeated measures dataset that contained six scientists and up to 40 of their most cited papers and how many times each of those papers were cited per Google Scholar. Figure 3 shows an initial dive into how citations change along increasing index, or the next highest cited paper for that scientist. Overall, this relationship does not appear to be linear for most of the scientists with many top cited papers receiving a lion’s share of the citations for that author and a steep drop off to the following most cited papers.

**Figure 3**

A line graph of different colored lines

Description automatically generated

*Figure 3* – Line graph depicting 6 different scientists most cited papers. The lower the index, the higher that article was ranked in terms of number of citations. Each colored line represents an individual scientist. Some scientists had more articles published than others indicated by some lines stopping before 40.

Like the previous analysis with the Anscombe data, I started with a spline fit for each scientist. The results of the spline fit are summarized and visualized in Figure 4.

**Figure 4**

A graph of a graph of a curve

Description automatically generated with medium confidence

*Figure 4* – Each scientists’ top cited papers plotted as connected points with the fitted individual spline models on top in dark red.

Since this is repeated measures data, a proper analysis for the full dataset would be a non-linear mixed effects model. However, to start, I whittled down the data to just one scientist (scientist D). I fit a nonlinear exponential model (using the equation below) to their data. The results are summarized in Table 3.

**Equation 2**

*Equation 2* – Exponential equation used for the nonlinear model fits to the scientist’s citation count data. The A coefficient represents where the curve intercepts the y-axis, and the B coefficient can change the steepness/shallowness of the curve.

**Table 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | SE | t | p |
| A (intercept) | 794.00 | 64.300 | 12.30 | <.001 |
| B | -.25 | .025 | -9.92 | <.001 |

*Note*. Nonlinear exponential fit to the citation count of scientist D’s top papers. The A parameter is where the curve intersects the y-axis (index 0) and the B parameter reflects the steepness of the curve or as it is more commonly known, the decay rate.

The scientist’s fitted curve is plotted over their citation counts for their top papers in Figure 5. The curve is steep, modelling a large drop-off after the top couple of papers, which take up a large share of their total number of citations.

**Figure 5**

A graph of a number of dots

Description automatically generated with medium confidence

*Figure 5* – Fitted nonlinear curve to scientist D’s papers and the number of citations for each (represented by individual points). The error ribbons represent 95% C.I. calculated using the predictNLS command in the propagate library in R.

*Part Four*

Now, I will fit all of the scientists with nonlinear curve fits at the same time using the nlsList command from the nlme library in R. This function will use the same starting coefficient values for each of the scientists to fit individual curves to each. The equation I used was the same exponential equation used previously for scientist D (Equation 2). The starting values I chose were (A = 400 and B = -1). The results for each curve fit are summarized in Table 4. This model was able to capture the individual variation in citation curve shapes in terms of both intercepts (A) and decay rates (B). A visualization of each curve fit to their respective raw data is plotted in Figure 6. This figure shows how each of the different best fitting coefficients for each nonlinear model affect the shape and how well it fits the underlying data. Some scientists showed a more even distribution of fewer citations across their top 40 papers (B, F), while those with more overall citations showed that large peak and drop-off from their most cited papers.

**Table 4**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | SE | t | p |
| Scientist A |  |  |  |  |
| A (intercept) | 1400.24 | 63.068 | 22.20 | <.001 |
| B | -.43 | .022 | -19.78 | <.001 |
| Scientist B |  |  |  |  |
| A (intercept) | 88.73 | 16.903 | 5.25 | <.001 |
| B | -.07 | .019 | -3.69 | <.001 |
| Scientist C |  |  |  |  |
| A (intercept) | 615.04 | 58.857 | 10.45 | <.001 |
| B | -.41 | .044 | -9.17 | <.001 |
| Scientist D |  |  |  |  |
| A (intercept) | 794.17 | 37.801 | 21.01 | <.001 |
| B | -.26 | .015 | -16.88 | <.001 |
| Scientist E |  |  |  |  |
| A (intercept) | 507.61 | 29.624 | 17.14 | <.001 |
| B | -.18 | .013 | -13.27 | <.001 |
| Scientist F |  |  |  |  |
| A (intercept) | 135.86 | 17.669 | 7.69 | <.001 |
| B | -.08 | .014 | -5.49 | <.001 |

*Note.* Coefficient values for each of the respective curve fits to individual scientists’ paper and citation count data. The change in parameters of the intercepts and decay rates reflect the model fitting these individual curves to the data.

**Figure 6**

A graph of a number of different colored lines

Description automatically generated with medium confidence

*Figure 6* – Plotted nonlinear curves over raw citation count data for each scientist (A-F) fitted with the nlsList command from the nlme library in R. Each scientist receives their own individual exponential curve fit, reflected with variations in the coefficients chosen to best fit their respective data.

*Part Five*

An exponential curve is not the only way to model this steep drop-off type of relationship seen in the citations of a scientist’s most cited papers. Another function that I considered was the power function (Equation 3).

**Equation 3**

*Equation 3* – Power function fit to the data using nonlinear regression. The A parameter is where the curve intercepts the y-axis, and the B parameter alters the shallowness/steepness of the curve pushing the bow of the curve away from the origin (0, 0).

The power function allows me to try a new relationship to fit the same data and compare that fit to the exponential curve fits to each scientist I’ve already done. So, I used the same nlsList command from the nlme library in R with starting values of A = 200 and B = 1, which seemed like good intermediate values between the steepest curves (like scientist A) and the flattest curves (like scientist F). The results of the power curve fits are summarized in Table 5. A visual depiction of the curve fits over the respective data are shown in Figure 7. The curves seem to model the data just as well as the exponential in some cases per visual inspection, but in some instances the curves seem to be fitting better. To get a true sense of which model was providing the better fit, I opted to use AIC to compare how well each model fits each scientist (Table 6). Five out of the 6 scientists’ data were better fit by the power curve than the exponential curve according to AIC. I think this could be due to the exponential curve having trouble unsticking from the axis, as it seems to lead to some issues finding a well-fitting line through the especially steep curves. Perhaps the power function is better at capturing this relationship with more sensitive parameters that allow for more nuanced curves.

**Table 5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | SE | t | p |
| Scientist A |  |  |  |  |
| A (intercept) | 964.36 | 20.602 | 46.810 | <.001 |
| B | 1.14 | .060 | 18.94 | <.001 |
| Scientist B |  |  |  |  |
| A (intercept) | 101.69 | 15.792 | 6.44 | <.001 |
| B | .19 | .061 | 3.11 | <.001 |
| Scientist C |  |  |  |  |
| A (intercept) | 436.21 | 20.401 | 21.38 | <.001 |
| B | .92 | .098 | 9.37 | <.001 |
| Scientist D |  |  |  |  |
| A (intercept) | 770.63 | 20.270 | 38.02 | <.001 |
| B | .82 | .048 | 17.06 | <.001 |
| Scientist E |  |  |  |  |
| A (intercept) | 505.42 | 19.420 | 26.03 | <.001 |
| B | .52 | .042 | 12.45 | <.001 |
| Scientist F |  |  |  |  |
| A (intercept) | 150.74 | 15.921 | 9.47 | <.001 |
| B | .20 | .043 | 4.58 | <.001 |

*Note*. Coefficient values for each of the respective curve fits to individual scientists’ paper and citation count data. The change in parameters of the intercepts and decay rates reflect the model fitting these individual curves to the data.

**Figure 7**

A graph of a function

Description automatically generated with medium confidence

*Figure 7* – Power function curve fits to the individual scientists’ (A-F) citation count data for their top papers. For some scientists (like D and E), it seems upon visual inspection that the curve is fitting the data better than the exponential curve for the same panels in Figure 6.

**Table 6**

|  |  |  |  |
| --- | --- | --- | --- |
| Scientist | AIC (power) | AIC (exponential) | AIC difference |
| A | 228 | 213 | 15 |
| B | 270 | 277 | -7 |
| C | 355 | 389 | -34 |
| D | 353 | 430 | -77 |
| E | 352 | 374 | -22 |
| F | 225 | 276 | -51 |

*Note.* AIC for each nonlinear curve fitted (power, exponential) to each scientist (A-F) and the AIC difference. Lower AIC indicates better fit for that model, and negative AIC differences indicate better fit for the power function specifically. AIC differences greater than 10 are strong evidence for better fit for that model. Five out of six models favor the power function compared to the exponential.

*Part Six*

Lastly, I used a spline fit within a mixed effects model using index as a fixed effect and one model with a random intercept of scientist and another model with a random slope of index by scientist as well as a random intercept. The splines were fit with a spline function (bs command within the splines library) within the lmer command from the lme4 library in R. The results are presented in Figure 8, with separate dashed and solid lines representing the random intercept and random slope curves for each individual scientist. Each parameter in an exponential or curve-based model like the spline is dependent upon one another. If one parameter is held constant, it could prevent the curve from contorting into different shapes that are appropriate for the data. In Figure 8, this is shown with the solid lines of the intercept only model that does not allow for variation in the slope or decay-rate-like parameter resulting in pretty much identical curves for each of the individual scientists. The random slope model was able to allow for more variation in the model parameters between subjects, and thus was able to capture more individual variation in curve shape between them.

**Figure 8**

A graph of different colored lines

Description automatically generated

*Figure 8* – Each scientists’ citation data were plotted at once within a mixed effects model. The solid lines represent the random intercept only model and the dashed lines represent the random slope model.

*General Discussion*

After going through this exercise, it seems like scientists who have a lot of citations in general are likely to have more of those citations within a handful of extremely influential papers. This matches what I’ve heard about publishing in other areas like novels and music, where most of the sales or airtime goes to a handful of prolific and influential individuals, while most novelists or musicians do not see a whole lot of success. It seems that this could be the same for publishing psychologists as evidenced by the well-fitting power and exponential curves to these 6 individuals.