STATISTICS TEACHER
SUPPORTING THE TEACHING AND LEARNING OF STATISTICS
STEW STatistics Education Web
online journal of K-12 statistics lesson plans

NONLINEAR MODELING: SOMETHING FISHY

Douglas Whitaker, Mount Saint Vincent University
Published: April 2021

Overview of Lesson
In this lesson students explore nonlinear regression models to explain fish weight using fish length, using both transformation of the response variable and polynomial regression. Geometric interpretations of variables are leveraged to suggest nonlinear models to fit. The intention of this lesson is for students to perform two or three linear regression analyses that feel like others that they have done before: the difference is that they draw on prior knowledge of geometric/physical relationships to suggest a modification to the first analysis to improve it. Because most of the nonlinear models considered in this lesson have only a single predictor variable, students' familiarity with simple linear regression can be extended to nonlinear modeling. If students are familiar with multiple linear regression, then two additional polynomial regression models can be included.

Type of Data
- Two quantitative variables
- Static dataset provided by lesson plan authors

Learning Objectives
- Students will propose and evaluate nonlinear models (transformation of response variable and/or polynomial regression)

Audience
- Students in a course that has a unit/chapter on linear regression in a statistical context, likely students in grades 9-14.
- *Prerequisites:* Prior to this lesson, students should have experience with graphing polynomials, radicals (square roots and cube roots), simple linear regression, and assessing linear regression fit (at least heuristically).

Time Required
60 to 90 minutes, depending on the how familiar students are with regression modeling and the software you use.

Technology and Other Materials
- *Technology:* Data analysis software appropriate for conducting simple linear regression. To accommodate a variety of software tools and modeling approaches, several versions of the data files are included. These are described later. If no technology for analysis is available, sample computer output is included that can instead be provided to students.
- "Low-tech" materials required: none, though depending on the students, it may be beneficial to have base-10 blocks or other cubes available to remind them of relationships among length, area, and volume, and to motivate the relationship between volume and weight.

Lesson Plan

The goal of this lesson is to develop an appropriate model for fish weight using fish length as a predictor. This lesson is appropriate to use after students are familiar with analyzing bivariate datasets using simple linear regression and heuristically assessing model fit. Students will begin by analyzing a dataset using simple linear regression, assessing its fit, and determining that a model of the form $\hat{y} = b_0 + b_1 x$ is NOT appropriate. Then, rather than stopping, a short review of geometric concepts is conducted. The idea of this review is for students to recognize two things: 1) a fish's weight is closely related to the fish's volume and 2) the units of volume are cubic millimeters. Together, this suggests we should explain fish weight with fish volume – a variable we don't have – but re-expressing our explanatory variable so that it has the units of volume might be a productive path forward.

Note: Throughout the lesson, the term *weight* is used for what is more properly described as mass. The original dataset uses the term *weight* but measures the fish using metric units for mass. The distinction is irrelevant to the lesson, but it would be reasonable to make the changes throughout the lesson if the distinction matters to you.

Description of Dataset

Scientists are interested in monitoring the health of trout perch in the Oil Sands Region of Canada.[1,2] As part of a larger study, fish were collected from the Athabasca River and Peace River, and several characteristics were measured, including the *Weight* of each fish (in grams) and the *Length* of each fish (in millimeters). The dataset used in this analysis includes data from 2088 trout perch. The first six rows of this dataset are shown to illustrate what the file looks like.

FishID	Length	Weight
4394	72	4.349
3302	54	1.916
4780	61	2.592
13184	58	2.115
6	73	4.118
13321	93	9.469

Initial Analysis (Simple Linear Regression)

Ask the students to examine the relationship between fish *Weight* and fish *Length* using simple linear regression and to assess the appropriateness of the model. This can be done using any tools/technology that students have used in their previous experiences with regression.

Students work in groups to analyze the dataset using simple linear regression and assess the model fit. Students record their decisions, results, and conclusions in the *Simple Linear Regression* section of the student handout. The questions they are asked to answer are:

- What regression model will you fit to explain Weight using Length as a predictor?
- What is the regression equation[3]?

---

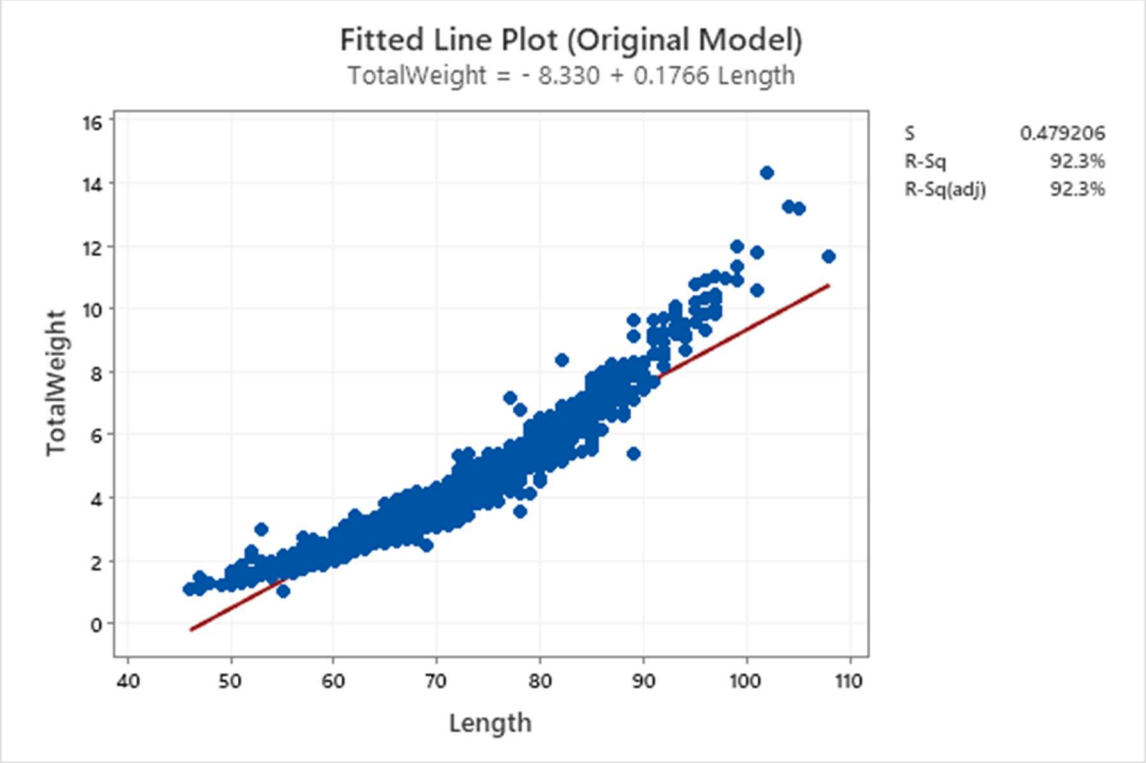
[1] Environment and Climate Change Data Canada http://data.ec.gc.ca/data/substances/monitor/fish-health-toxicology-contaminants-oil-sands-region/wild-fish-health-oil-sands-region/
Data License: Open Government Licence - Canada

[2] Trout perch illustration by Ellen Edmonson and Hugh Chrisp - http://pond.dnr.cornell.edu/nyfish/fish.html, Public Domain, https://commons.wikimedia.org/w/index.php?curid=4974275

[3] Feel free to substitute the term *prediction equation* or *fitted model* depending on what has been used in class.

- What is the predicted (fitted) value for the fish with ID 4394? (See data excerpt on pg. 1)
- What is the residual for the fish with ID 4394?
- Interpret the slope.
- Interpret the R^2 value.
- Examine the residual plots and comment on any potential problems.

Depending on the type of statistics course this lesson is being used in, students may have informal or formal ways for assessing model fit. At a basic level, a scatterplot exhibits a clear curved (nonlinear) relationship between length and weight, as seen in this graph:



Note that the weights of the shortest and longest fish tend to be *underestimated* by the regression model (because the regression line is below the points for the smallest and largest values of Length); the weights of average length fish tend to be *overestimated* by the model, though this is far less clear in this graph because of how many observations are in the dataset.[4]

This can more clearly be seen in a *Residuals vs. Fits* graph.[5] This is a type of residual plot (sometimes called a residual diagnostic plot).[6] While many software packages include this as a standard regression

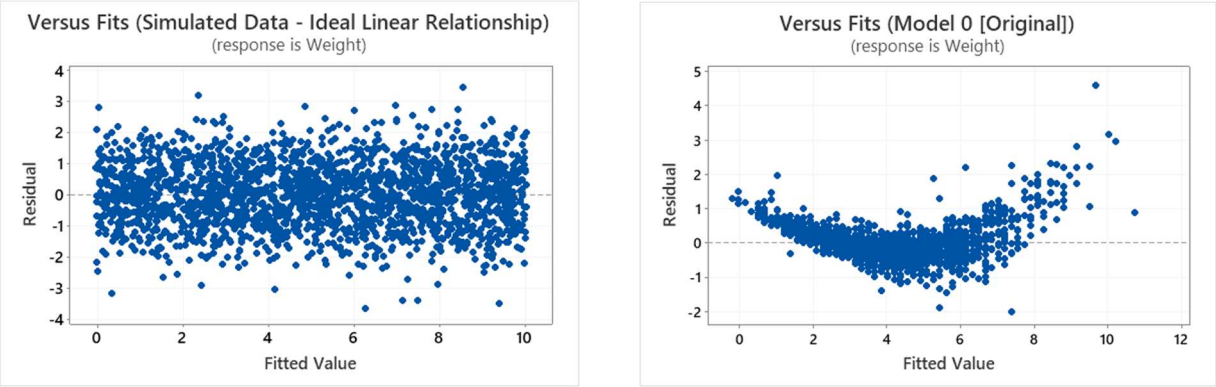
[4] See the Reflections and Additional Recommendations section at the end of the lesson plan for an extension to the lesson that focuses on the underestimation/overestimation for certain fish lengths.

[5] For simple linear regression, a scatterplot of the data with a fitted regression line and a *Residuals vs. Fits* graph can reveal essentially the same insights. The real advantage of residual plots is when more sophisticated models are used that cannot be easily visualized directly. Depending on what software you and your students are using, fitting models E and F below may result in difficulties visualizing the model fit. For these models, a *Residuals vs. Fits* graph is powerful tool for gaining insight about the model fit. Sometimes these graphs are referred to by similar names, such as *Versus Fits* (the term used in the Minitab output).

[6] See note at the end of the lesson about other residual plots.

plot, they can be constructed manually. To construct a *Residuals vs. Fits* graph, the fitted (predicted) value and residual must be calculated for each observation – a tedious task best-suited for computers. Then, a scatterplot is made with the residuals on the Y-axis and the fitted values on the X-axis; a horizontal line at $y = 0$ is usually added.

The ideal *Residuals vs. Fits* graph should have a horizontal scattering of points with no discernible pattern, with points approximately equally above and below the horizontal line throughout the entire range of the fitted values; a *Residuals vs. Fits* graph illustrating an ideal relationship using simulated data (randomly generated so that it has a linear relationship) is shown below. The *Residuals vs. Fits* graph for the original simple linear regression model differs from the ideal in two ways. First, there is a clear curved pattern, indicating the same nonlinearity observed in the scatterplot. Second, the spread of the residuals for low fitted values is smaller than the spread of the residuals for large fitted values – an issue known as non-constant variance.



Two *Residuals vs. Fits* graphs: on the left, an example of one for data with an ideal linear relationship (simulated data); on the right, the graph created using the residuals for Model 0.

Students can share their results and come to consensus about the appropriateness of the model in a manner that is typical for your class. Once students have analyzed the data using simple linear regression and determined that a linear model is likely NOT the most appropriate model to use, the lesson continues to the next stage.

Activating Prior Knowledge and Planning

The purpose of this stage is to activate students' prior knowledge and determine a path forward in the analysis. Students continue in their same groupings to answer the questions in the *Background* section of the student handout. These questions ask them to sketch the shapes of simple polynomial equations ($y = x$, $y = x^2$, and $y = x^3$) and think of a rectangular prism as a simple approximation for the shape of a fish (to motivate the polynomial relationships used later in the modeling).

There are two distinct paths forward for students that are both reasonable: transforming the *response* variable or including polynomial terms as *predictor* variables to address the curvature (polynomial regression). There are subtle differences in the approaches, but both can result in an appropriate model. In the following pages, transforming the response variable will be discussed first followed by polynomial regression. Depending on the amount of time available, you may wish to guide the students through only one of these possibilities – transformation of the response variable is the clear candidate for this as it may seem simpler and more accessible to students.

When looking at the shape of the scatterplot, it is reasonable for students to think the relationship is described by $y = x^2$ as it resembles half of a parabola. It is also reasonable for students to think that $y = x^3$ is appropriate because it resembles part of the graph of a cubic function. These forms might be more familiar to students and suggest polynomial regression. However, students may also recognize that $\sqrt{y} = x$ or $\sqrt[3]{y} = x$ might similarly be reasonable ways of expressing the relationship they see – these forms suggest a transformation of the response variable. Students should be encouraged to pursue either a 2nd- or 3rd-degree polynomial model or a transformation using square roots or cube roots. If there isn't sufficient variety in the models that students select on their own, consider assigning certain models to groups of students. Ensuring that diverse models are covered (e.g., Models A, B, C, and D shown below) will lead to a richer discussion at the final stage of the lesson when students compare their results with classmates.

Students should record the model they plan to fit in the New Model section of their student handout. In summary, the following models are all reasonable depending on the patterns that students identify.

Label	Model	Description
A	$\sqrt{\widehat{Weight}} = b_0 + b_1(Length)$	Simple Linear Regression Model with Square Root Transformation
B	$\sqrt[3]{\widehat{Weight}} = b_0 + b_1(Length)$	Simple Linear Regression Model with Cube Root Transformation
C	$\widehat{Weight} = b_0 + b_1(Length)^2$	Quadratic Regression Model, without lower-order terms *or* Simple Linear Regression Model with Square Transformation for the Predictor
D	$\widehat{Weight} = b_0 + b_1(Length)^3$	Cubic Regression Model, without lower-order terms *or* Simple Linear Regression Model with Cube Transformation for the Predictor
E	$\widehat{Weight} = b_0 + b_1(Length) + b_2(Length)^2$	Quadratic Regression Model
F	$\widehat{Weight} = b_0 + b_1(Length) + b_2(Length)^2 + b_3(Length)^3$	Cubic Regression Model

For the purposes of this lesson, fitting models of the form $\widehat{Weight} = b_0 + b_1(Length)^3$ is reasonable. However, in polynomial regression, it is typical to include all lower-order terms in a model. That is, if a 3rd-degree term is in a model, it should also include the 2nd-order and 1st-order terms, so: $\widehat{Weight} = b_0 + b_1(Length) + b_2(Length)^2 + b_3(Length)^3$. There are different schools of thought around this practice, but nuanced interpretations of the coefficients are easier when the lower-order terms are included; when they are not included, the interpretation of the coefficient of the highest-order term also must account for the lower-order components. If the course you are teaching does not cover multiple linear regression – or does not cover multiple linear regression until much later – this likely does not need to be mentioned.

Any polynomials of higher degree than this are probably too complicated to consider in this analysis. There is some evidence that a 4th-degree model would not be unreasonable, but the value added to this model by the additional complexity is likely not practically useful. A 3rd-degree model works well and is more interpretable than a 4th-degree model. Similarly, a quartic root transformation is likely unnecessarily complex given the real-world relationship motivating this dataset. If, say, one group wants to pursue something of this complexity, that should be fine – the model can then be compared in the next phase of the lesson and will likely not be preferred to one of the other models.

Note: if the software you are using does not support multiple linear regression or polynomial regression, then models E and F cannot be fit. Instead, students pursuing an approach consistent with polynomial regression (i.e., focused on making adjustments to the predictors) should be encouraged to fit models C or D, respectively. If using TI-series calculators (or a similar technology), note that the order of the fish observations in the datasets has been randomized: it would be reasonable to have students manually enter the first, say, 10, 20, or 30 observations into lists and then use calculator's functions to make appropriate transformations and perform the regressions. Models A, B, C, and D can all be fit in this way. (While 10 observations may minimally work, a curved pattern is apparent when juxtaposed with a fitted regression line using the first 30 observations.)

Fitting and Evaluating Nonlinear Models

Students now put their plan (developed in the second stage) into action. The logistics of how students enact their plan depends on the software tool used and the technical abilities of the students. A few possibilities are described below. During this stage, you will support the students as they use software to model and as they answer the questions in the *New Model* section of the student handout. Depending on the software available, your students' abilities, and your pedagogical goals, you may consider providing students with copies of computer output that has already been created (included as a supplement to this lesson).

For students comfortable with both linear regression and a full-featured statistical software package (e.g. Minitab, SPSS, R), students may be expected to fit their model directly from the raw data using the regression modeling and/or transformation tools built-in to the software. For students with less experience with fitting models, you may wish to suggest these tasks to them (presented in the list below in the approximate order they might be completed):

- Create new columns (new variables) corresponding to the transformed variables they plan to analyze (e.g., a column for $\sqrt{Weight}$ or $Length^3$).
- Make a scatterplot to explore bivariate relationships with these newly created variables.
- Perform a linear regression using these newly created variables.
- To make a residual plot:
    - Create a column of predicted fish lengths based on the regression equation for the new model.
    - Create a column of residuals by subtracting the predicted fish lengths from the actual fish lengths.
    - Make a scatterplot with residuals on the y-axis and predicted values on the x-axis.

If students have less experience with statistical software and/or linear modeling, you may wish to provide them with a dataset that already includes the appropriate calculated variables (included with

this lesson). If providing students with these pre-calculated variables, this file should not be shared with them until *after* they have determined which model they want to fit. Otherwise, seeing the calculated variables included in the dataset may suggest to them a path forward – the path forward should be suggested by the patterns they observe rather than the ease of fitting the model.

Once students have fit their new regression model, they answer questions about the new analysis in the *New Model* section of the student handout. The questions are largely the same as the ones they were asked to answer about the original simple linear regression model.

Note: Depending on which model students chose to fit, interpreting the value of the fitted slope might be more difficult or impossible; depending on your pedagogical goals, you may wish to omit the question asking students to interpret the slope of their alternative model (perhaps just sharing correct interpretations with them instead to illustrate the complexity).

- Propose a new model to explain Weight using Length as a predictor. Why did you choose this?
    - This was answered in the previous stage of the lesson.
- What is the regression equation?
- What is the predicted (fitted) value for the fish with ID 4394? (See data excerpt on pg. 1)
- What is the residual for the fish with ID 4394?
- Interpret the slope.
- Interpret the R^2 value.
- Examine the residual plots and comment on any potential problems.
- Overall, does this model appear to be a good fit for the data?

Then students compare the new model to the simple linear regression model in the *Model Comparison* section of the student handout.

- Comparison of their fitted model with the original model:
    - Which model has a higher R^2 value? What does this mean?
    - For which model is the fitted slope easier to interpret?
    - Compare the residual plots for the two models and comment on the similarities and differences.

In statistical practice, model fitting is often an iterative process where exploratory data analysis and theoretical knowledge motivate models. These models are fit, examined, and compared, which in turn motivates new models that may be considered. In this lesson, students engage in a limited facsimile of model building: an initial model is considered, an alternative model is proposed based on an examination of the initial model fit and theoretical considerations (i.e., the relationship between length, volume, and weight), and that new model is fit and examined. As a class, several models will be considered, but it would be reasonable to expect a single analyst to generate all of these models – and more – as part of the model fitting process.

Group Discussion to Compare Models

After students have finished fitting their model, evaluating it, and comparing it to the original model, you should facilitate a group discussion where each group shares its model and their findings. For each of the models that students might fit, the R^2 value will be higher than the original model. This suggests

that pursuing a nonlinear modeling approach improves the model fit. However, improvement in R^2 alone is not enough to adjudicate which model is best.[7] Students should also look to the residual plots for improvement in the curvature (and non-constant variance if this was noted earlier). Depending on which model students fit, they may see an improvement in the residual plots but still note that problems exist. (Sometimes this is the best we can hope for!)

Students should then discuss which model (or models) among all the fitted models is best. This section of the lesson plan offers a brief overview of the model comparisons. The following sections provide the fitted values and residuals, the interpretation of the slopes, and evaluation of the models in much more detail.

Table with the regression equation and R^2 value for all models considered.

Label	Description	Regression Equation	R^2	*Residuals vs. Fits* Plot
0	Simple Linear Regression (original model)	Weight = - 8.330 + 0.1766 Length	.9231	Clear curvature, potential non-constant variance
A	Simple Linear Regression Model with Square Root Transformation	Weight^(1/2) = - 0.9896 + 0.04222 Length	.9601	Mild potential curvature, no apparent non-constant variance
B	Simple Linear Regression Model with Cube Root Transformation	Weight^(1/3) = 0.007328 + 0.02221 Length	.9637	No apparent curvature, no apparent non-constant variance
C	Quadratic Regression Model, without lower-order terms	Weight = - 2.151 + 0.001239 Length^2	.9506	Clear curvature (though improved over the original mode), potential non-constant variance
D	Cubic Regression Model, without lower-order terms	Weight = - 0.02667 + 0.000011 Length^3	.9626	No apparent curvature, potential non-constant variance
E	Quadratic Regression Model	Weight = 5.766 - 0.22211 Length + 0.002768 Length^2	.9628	No apparent curvature, potential non-constant variance
F	Cubic Regression Model	Weight = 0.80 - 0.0133 Length - 0.000108 Length^2 + 0.000013 Length^3	.9630	No apparent curvature, potential non-constant variance

Note that all models exhibit non-normality of the residuals if one examines the Normal Probability Plot, though this is not central to the lesson.

The cube root transformation model (model B) is likely the best model of the ones considered. Not only does it have the highest R^2 value, it more importantly addresses the curvature in the residual plot

[7] As more predictors are added to a linear regression model, the R^2 value will *always* increase or stay the same – never decrease. This is because $R^2 = 1 - \frac{SSError}{SSTotal}$: $SSTotal$ is the same for all models that use the same set of Y values, and $SSError$ can only become smaller (or stay the same) as additional predictors are included in a model.

noted in the original model, the non-constant variance noted in the original model, and has a relatively straightforward interpretation. Other comparisons of similar pairs of models are:

- Between models A and B, model B would be preferred because of its improvement in the residual plots.

- Model C is likely not defensible relative to any of the others: while it represents and improvement over the original model, there are still clear problems in the residual plots. Model D would be preferred to C because of the improvement in the residual plots.

- Models E and F are quite similar, and – based on the model comparisons examined in this lesson – neither is definitively better than the other. Still, if a single model must be chosen, model F might be preferred because it is consistent with the real-world relationship that weight is closely related to volume, which is three-dimensional property.

Students should summarize the group discussion in the *Overall Comparison* section of the student handout.

- Overall comparison:
    - Out of all of the models, which model do you prefer?
    - Why?

Fitted values and residuals

The student handout asks students to make a prediction and calculate a residual only for fish ID 4394 (the first row in the included image of the 6 rows of the dataset). You could modify this value easily to be any of the other IDs included in the image of the first 6 rows (or even a different value from the dataset not pictured). With how many possible answers there are for this (considering the models students may fit), this may not be expedient for grading. To facilitate grading, the table below gives the fitted value and residual for the first six observations for the original model and models A-F. (These values were obtained using a computer: depending on how and when students round, the values they obtain may be slightly different.)

Table showing the fitted value and residual for the first six observations in the dataset for each model considered in this lesson.

	FishID	4394	3302	4780	13184	6	13321
	Length	72	54	61	58	73	93
	Weight	4.349	1.916	2.592	2.115	4.118	9.469
Model 0 (Original)	Fitted Value	4.384	1.205	2.441	1.912	4.560	8.092
	Residual	-0.035	0.711	0.151	0.203	-0.442	1.377
Model A	Fitted Value	2.050	1.290	1.586	1.459	2.092	2.937
	Residual	0.035	0.094	0.024	-0.005	-0.063	0.140
Model B	Fitted Value	1.606	1.207	1.362	1.295	1.629	2.073
	Residual	0.026	0.035	0.012	-0.012	-0.026	0.043

Model C	Fitted Value	4.272	1.462	2.460	2.017	4.452	8.566
	Residual	0.077	0.454	0.132	0.098	-0.334	0.903
Model D	Fitted Value	4.155	1.738	2.517	2.159	4.332	8.986
	Residual	0.194	0.178	0.075	-0.044	-0.214	0.483
Model E	Fitted Value	4.126	1.845	2.519	2.197	4.305	9.054
	Residual	0.223	0.071	0.073	-0.082	-0.187	0.415
Model F	Fitted Value	4.127	1.810	2.533	2.197	4.303	9.071
	Residual	0.222	0.106	0.059	-0.082	-0.185	0.398

Comparing the analyses can take several forms: the appropriateness of the model fit and/or how well the explanatory variable accounts for variability in the response variable (i.e. comparing R^2 values).

Interpretation of slope

On their worksheet, students are asked to interpret the fitted slope for the original model and their new model. The interpretation of the slope for the original model may not be fully appropriate (considering problems were identified), but a standard interpretation[8] might be:

> *For every 1-millimeter increase in fish length, we expect a 0.1766-gram increase in fish weight, on average.*

Because these units are relatively small, it might make sense to adjust the interpretation, such as:

> *For every 10-millimeter increase in fish length, we expect a 1.766-gram increase in fish weight, on average.*

Depending on which model students chose to fit, interpreting the value of the fitted slope might be more difficult or impossible; depending on your pedagogical goals, you may wish to omit the question asking students to interpret the slope of their alternative model (perhaps just sharing correct interpretations with them instead to illustrate the complexity). A full interpretation for the fitted slope for each alternative model follows.

A. Because this is a simple linear regression model, the interpretation is more straightforward – but there is still a challenge. A reasonable interpretation might be:

   For every 1-millimeter increase in fish length, we expect a 0.04222-$\sqrt{gram}$ increase in fish $\sqrt{weight}$, on average.

   In this model, the response variable is *not* $Weight$; the response variable is $\sqrt{Weight}$ and the interpretation should reflect that. The interpretation is more straightforward than for models A and C because the explanatory variable is still simply $Length$, but this model does not establish

[8] Adjust the interpretation as needed to include aspects you value, such as restricting the interpretation to only apply to trout perch collected from the Athabasca River and the Peace River.

a linear relationship between $Length$ and $Weight$ – it instead establishes a linear relationship between $Length$ and $\sqrt{Weight}$, which is a nonlinear relationship.

B. As in model A, this is a relatively straightforward interpretation but still challenging because of its nonlinear nature. A reasonable interpretation might be:

For every 1-millimeter increase in fish length, we expect a 0.02221-$\sqrt[3]{gram}$ *increase in fish* $\sqrt[3]{weight}$*, on average.*

In this model, the response variable is *not* $Weight$; the response variable is $\sqrt[3]{Weight}$ and the interpretation should reflect that. The interpretation is more straightforward than for models A and C because the explanatory variable is still simply $Length$, but this model does not establish a linear relationship between $Length$ and $Weight$ – it instead establishes a linear relationship between $Length$ and $\sqrt[3]{Weight}$, which is a nonlinear relationship.

C. Because this model uses $Length^2$ as the only predictor, the fitted slope value of 0.001239 is the expected increase in $Weight$ (in grams) for a 1-unit increase in $Length^2$. Of course, $(1\ mm)^2 = 1\ mm^2$, so the interpretation can be somewhat simplified. However, if students which to modify the interpretation as was done above (i.e. describing a 10-millimeter increase in $Length$), this becomes much more challenging. It is also worth noting that this does not refer to a 1-millimeter^2 increase in the surface area of the fish: only the *length* is being measured, and surface area is a distinct property which was not measured. The slope also incorporates the change due to the linear term that was omitted, further complicating the interpretation.

D. As in model C, the interpretation is complicated but possible. Because this model uses $Length^3$ as the only predictor, the fitted slope value of 0.000011 is the expected increase in $Weight$ (in grams) for a 1-unit increase in $Length^3$. Of course, $(1\ mm)^3 = 1\ mm^3$, so the interpretation can be somewhat simplified. However, if students which to modify the interpretation as was done above (i.e. describing a 10-millimeter increase in $Length$), this becomes much more challenging. It is also worth noting that this does not refer to a 1-millimeter^3 increase in the volume of the fish: only the *length* is being measured, and volume is a distinct property which was not measured. The slope also incorporates the change due to the linear and quadratic terms that were omitted, further complicating the interpretation.

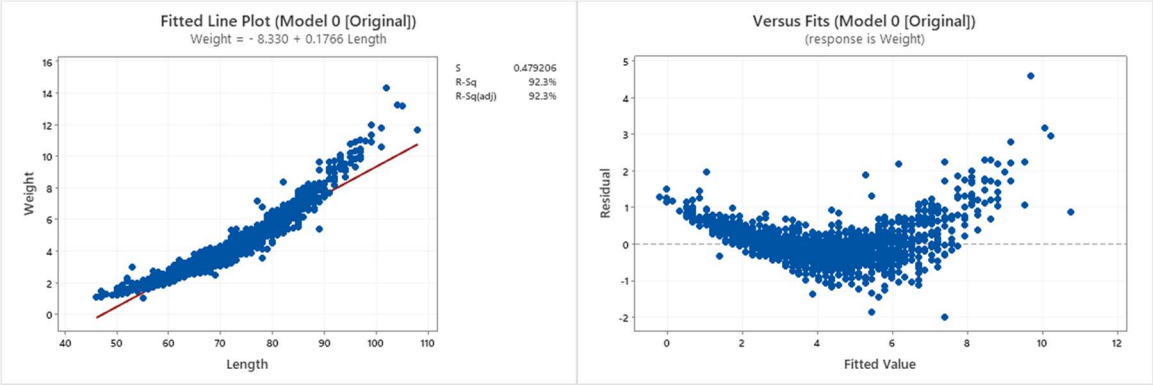
E. Because this model includes both $Length$ and $Length^2$ as predictors, there is not a single slope to interpret. This makes sense: interpreting the slope in linear regression describes the *linear* change we expect in the response for a change in the explanatory. In this model, we are fitting a nonlinear model and cannot simply interpret it as we would a linear model.

F. As in model E, the slope cannot be interpreted for this model because there is not a single slope: there are three slopes for this nonlinear model. Interpreting the slope in linear regression describes the *linear* change we expect in the response for a change in the explanatory. In this model, we are fitting a nonlinear model and cannot simply interpret it as we would a linear model.

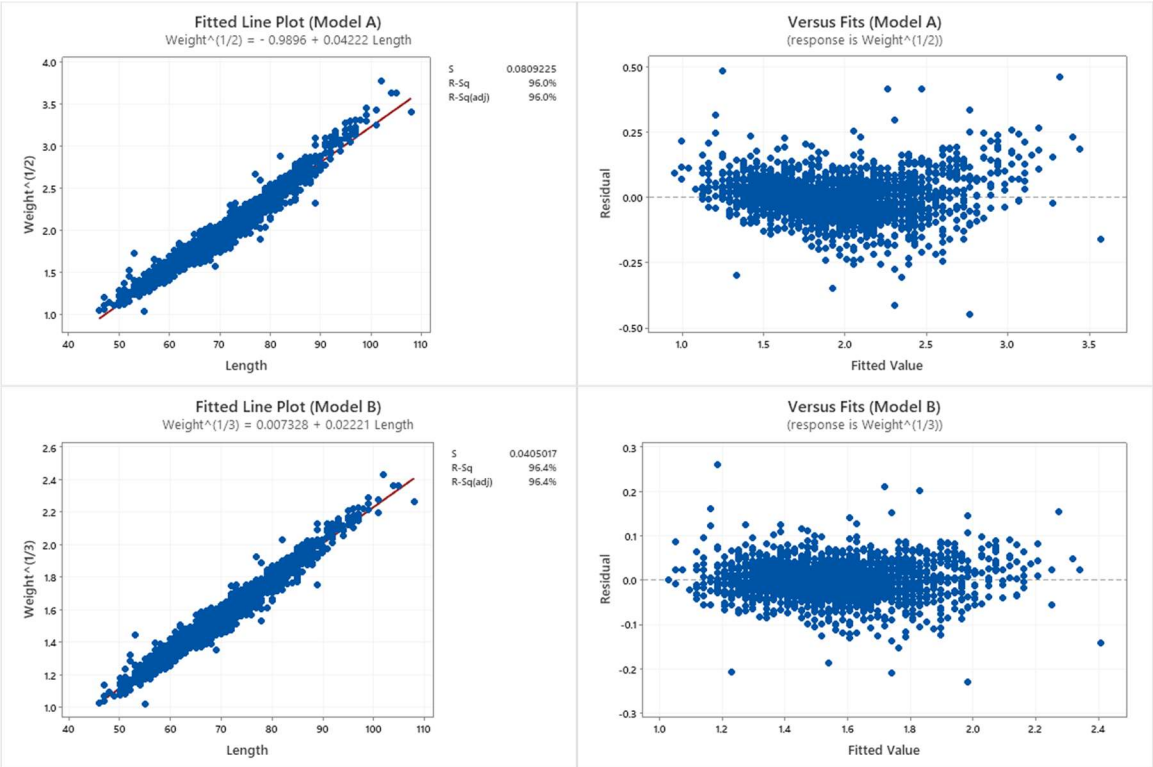
If students fit a model using either a 4th-degree polynomial or a quartic root transformation, the challenges in interpretation are similar to the above.

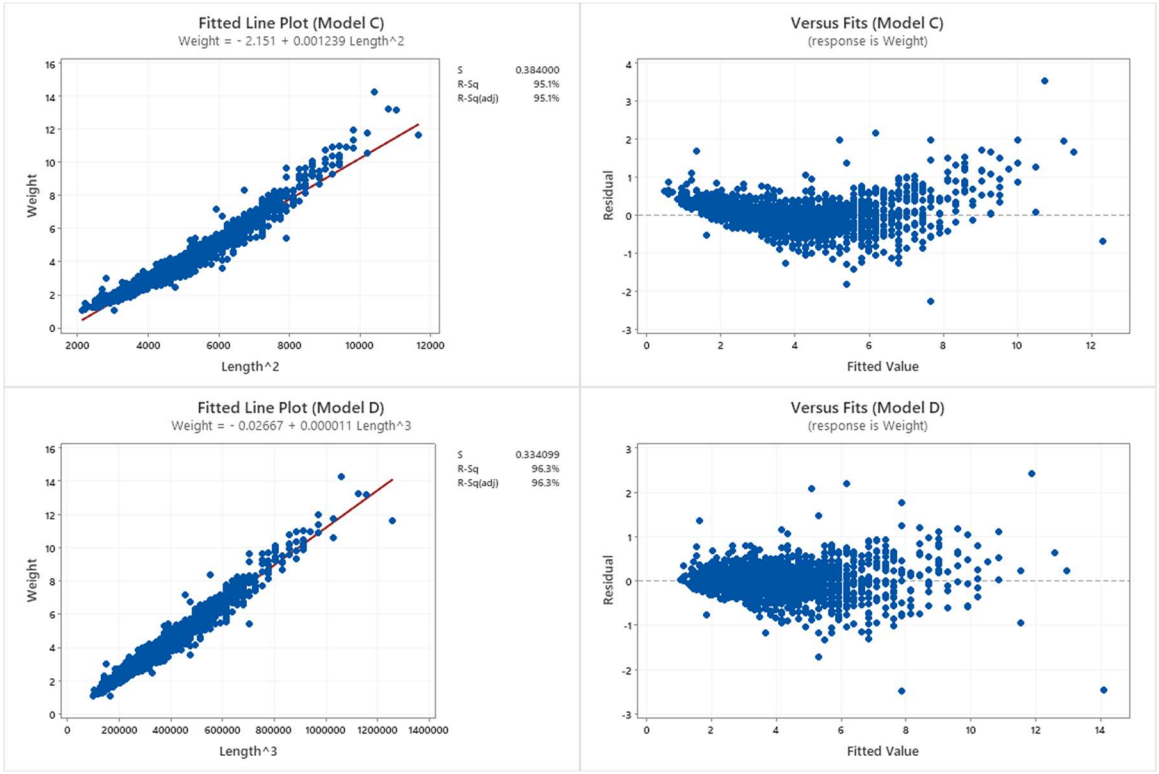
Evaluation of models

The scatterplot with the regression fit and the *Residuals vs. Fits* graph for each model are shown below. While each model (A-F) is an improvement over Model 0, Model B is perhaps the best among the models considered based on the *Residuals vs. Fits* graph (though not overwhelmingly better than any other model). A note at the end of this document includes additional comments about residual plots.

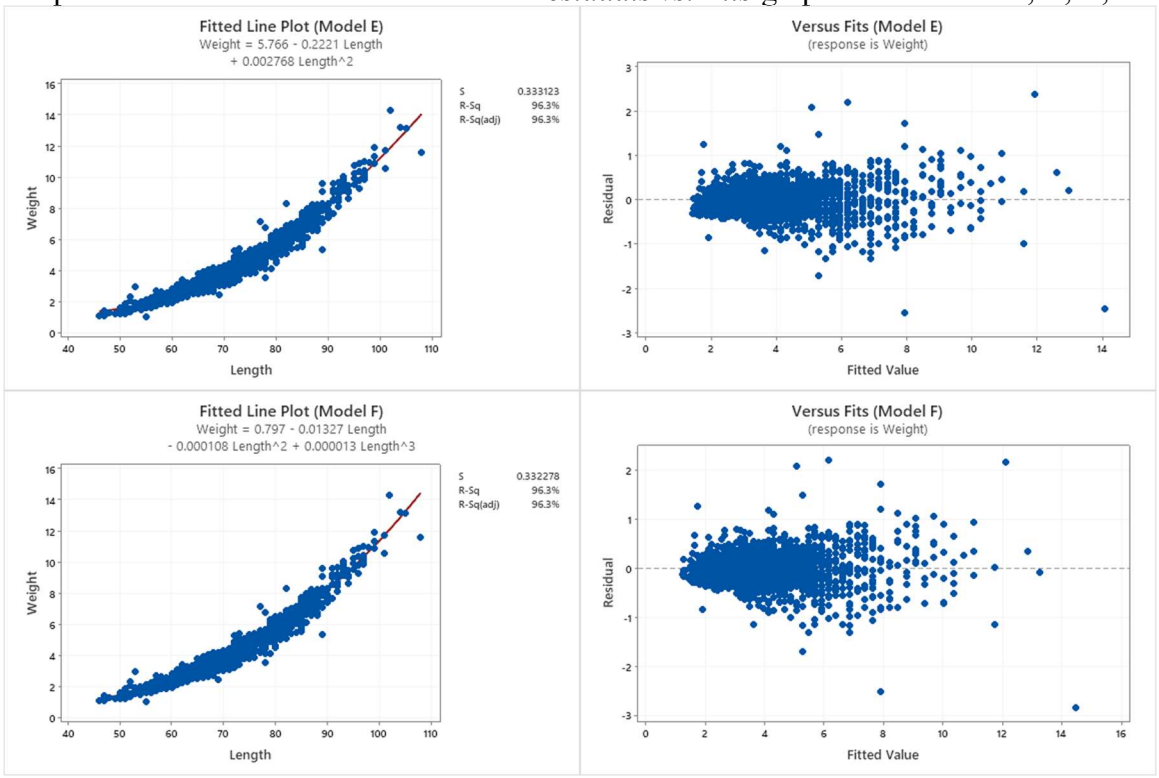


Scatterplot with fitted models and residual *Residuals vs. Fits* graph for Model 0 (Original)





Scatterplots with fitted models and residual *Residuals vs. Fits* graphs for Models A, B, C, and D



Scatterplots with fitted models and residual *Residuals vs. Fits* graphs for Models E and F

Attached Materials

The following files are included:

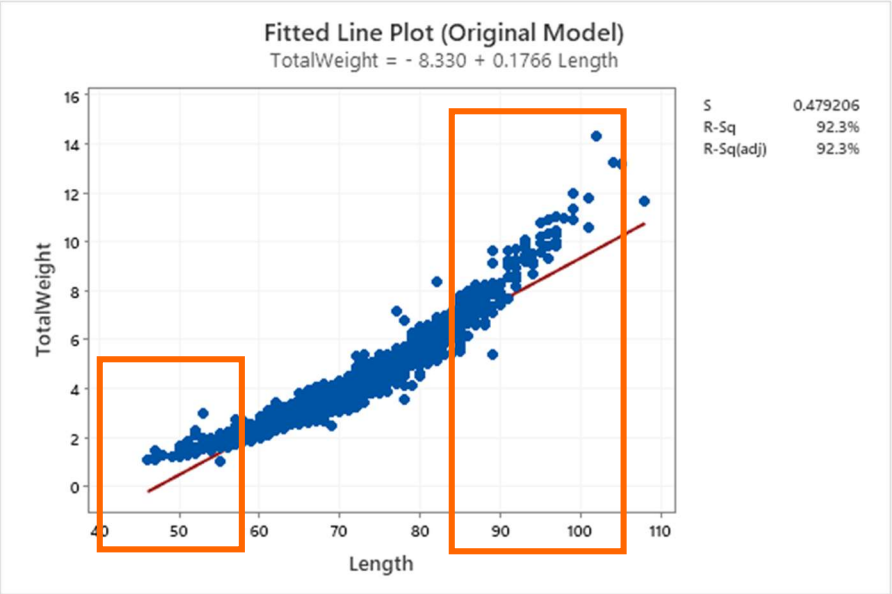
- Handouts
  - Student handout (Word file)
  - Selected computer output from the analyses for the Trout Perch data that can be used if students do not have access to appropriate technology (Word file; included in the .zip file containing the data files)
- Data files
  - Trout Perch data, cleaned and appropriate for analysis (Minitab and .csv)
  - Trout Perch data, cleaned and appropriate for analysis, with calculated variables (Minitab and .csv)
  - An additional fish dataset from the same source (Slimy Sculpin) that exhibits the same relationships and can be used to create exam questions, assignments, etc.
  - The raw data files for both Trout Perch and Slimy Sculpin (as they were downloaded from the Environment and Climate Change Canada website).
- Technology guides
  - Minitab
    - A brief overview for using Minitab to perform the analyses in this lesson
    - A Minitab workbook file containing various analyses already performed for the Trout Perch Data.
  - CODAP
    - A brief overview for using CODAP to perform some of the analyses in this lesson
    - A CODAP file containing two of the models and residual plots as an example.
  - TI-83/84
    - A brief overview for using TI-83/84 calculators to perform the analyses in this lesson (with a subset of the data)

Reflections and Additional Recommendations

Thoughts on extensions

Students are asked to make a prediction and calculate a residual for a single observation using both Model 0 and the model that they propose. By itself, a single residual often has little value – even answering questions such as "Is this residual particularly large (or small)?" require consideration of the residual relative to others. This discussion of residuals can be extended to more precisely articulate the shortcomings of Model 0: you can ask students to explain if the original model will systematically overestimate or underestimate the weights of any types of fish. As indicated by the rectangles below, the shortest and longest fish will tend to have their weights *underestimated* by the original regression model. There is a corresponding tendency for fish with lengths closer to the average length to have their weights *overestimated*, though this is less clear from the graph due to the abundance of observations obscuring the line. Such a discussion might be prompted by assigning each group of

students a different observation (intentionally chosen because the model overestimates or underestimates it[9]) and identifying patterns among the set of residuals across groups.



This discussion could also be related to extrapolation: to what extent would Model 0 or any of the alternative models be appropriate for fish shorter than the 46 mm or longer than 108 mm (the minimum and maximum, respectively)? While extrapolation can be dangerous and should not be undertaken lightly, Model 0 would perform *even worse* for extrapolations, but the alternative models might be more cautiously applied for fish close to the lengths in our dataset (e.g., 40 mm and 110 mm) because we seem to have accounted for the salient curve in the data and we have a model that has some basis in a physical relationship. Still, consulting with someone who has expertise in fish biology would be advisable before using any of these models to make moderate extrapolations outside of a classroom setting.

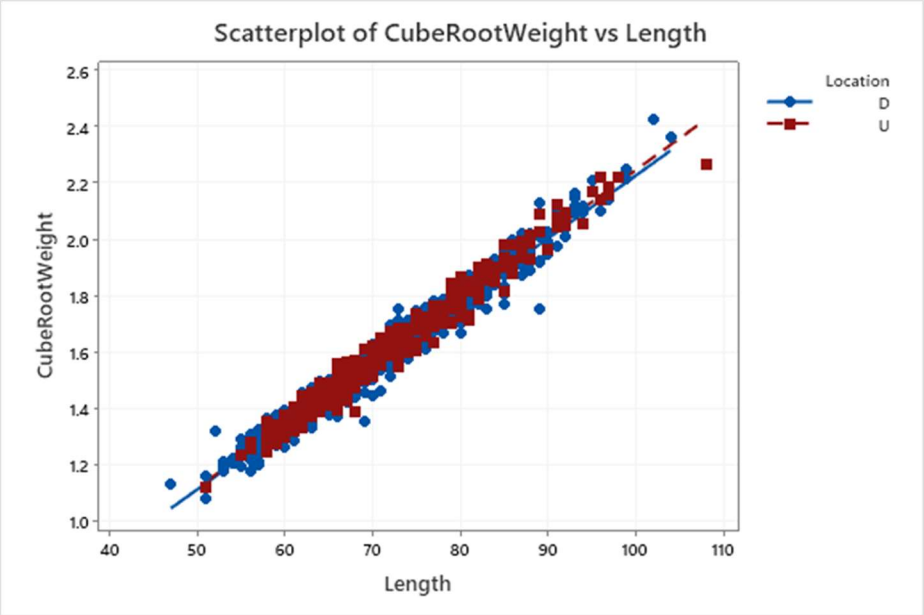
When using polynomial regression (and multiple linear regression in general), model over-fit is a substantial concern. There can be a tendency for some students to want to fit very complex models (e.g. degree n-1 polynomials) that can result in perfect fit for the dataset they are using but would be terrible models when applied to any other dataset. This suggests two extensions:

1. Discuss the behaviour of the fitted models *just outside* the values that are included in the dataset. Often, polynomial models will appear quite unrealistic for values even slightly outside the dataset.
2. Explore model validation techniques such as cross-validation. Each group could be given a slightly different dataset of randomly chosen fish. The groups could develop their models and then apply them to the remaining fish and calculate appropriate fit statistics. Many textbooks on linear regression include such a topic, including *STAT2* by Cannon et al. (2018).

[9] Some candidates for these observations are the fish with the following FishID: 3319, 3331, 3343, 3344 (long fish that are all underestimated); 3327, 4052, 4151, 13320 (mid-length fish that are all overestimated); 4, 3259, 4091, 4198 (short fish that are all underestimated). Note that because the FishID values are in a random order, it may be easiest to sort the dataset by FishID in a spreadsheet program (such as Microsoft Excel) to identify these observations. Again, these observations were intentionally chosen to illustrate the underestimation/overestimation pattern – there are, of course, observations that do not fit this pattern, such as 1196, 3165, 3179, 4282 (mid-length fish that are all *under*estimated).

This dataset was collected as part of a study to monitor fish health in the context of development in Canada's Oil Sands Region. Specifically, fish were sampled upstream from developed sites to serve as a reference for fish sampled downstream from developed sites that could exhibit environmental effects from the developed site. Facilitating a discussion with students about *how* a model developed in this lesson plan (e.g., Model B) could be used to determine if environmental impacts were occurring could connect this lesson back to the research context and motivate further statistical considerations. In the raw data included with this lesson, information about where fish were sampled is included (US = upstream; DS = downstream); the dataset used in this lesson includes fish from both upstream and downstream locations. Possible approaches students may consider are:

- Fitting a model and determining if new fish are much smaller (or larger) than predicted, leading to a discussion about residuals and unusual points
- Including the upstream/downstream information in the model as a categorical variable. For less experienced students, this might take the form of using colours to indicate upstream or downstream group membership on a graph and/or fitting two regression models and visually inspecting them for differences. See the graph below for an example. You could ask, "how much difference in regression coefficients is enough difference to determine if the developed sites seem to have an effect on fish?" to motivate inferential techniques with regression. More advanced students could use more formal approaches and possibly consider interaction terms.



Note that the there is a rather small coefficient for the categorical predictor variable, but the p-value is also rather small (about 0.001). This could motivate a discussion about statistical significance versus practical significance.[10]

[10] The term *statistical significance* is no longer recommended by many (e.g., Wasserstein et al., 2019). Such a small difference coefficient could be used as an example of when the term *significance* might not convey the intended message.

Thoughts on differentiation

Depending on where students are at in their learning experience, you may wish to provide some of the students with the dataset that contains the calculated variables already while other students may be asked to use software to calculate the variables themselves. Moreover, some groups could be focused on just one type of model (i.e. polynomial or transformation of the response) from the beginning, which would provide more structure. Due to the complexity of the interpretations of the slope for various models, providing some structure to this (e.g., providing students with a template and have them fill in blanks) would be another way to support students with different learning needs and could be combined with other modifications.

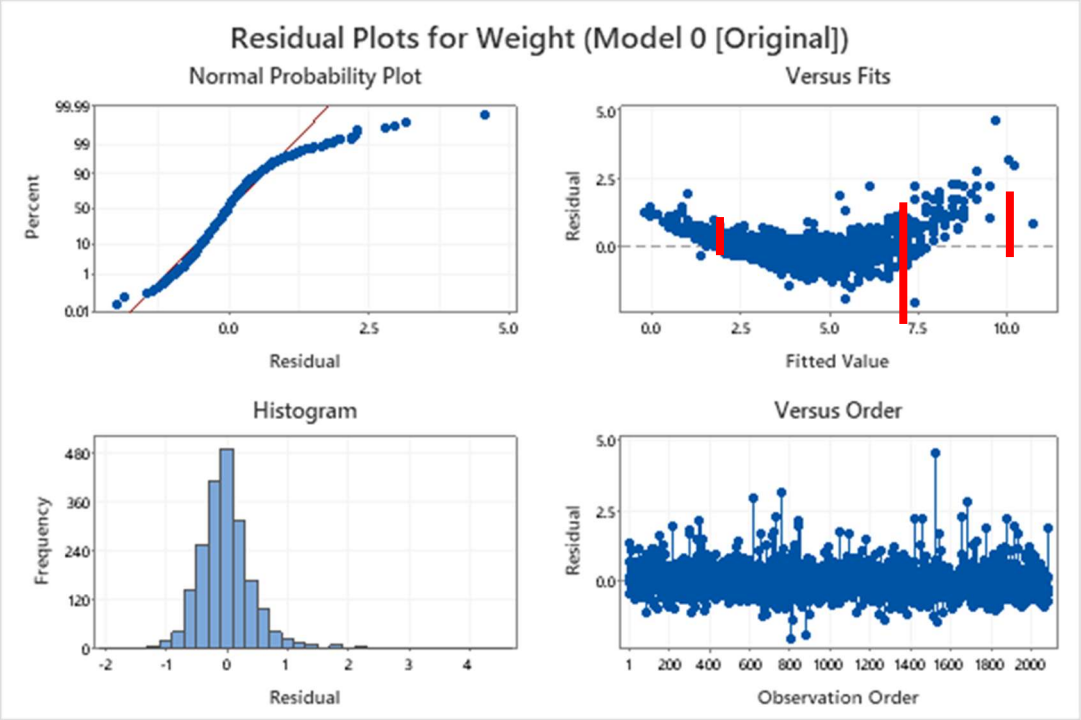
Thoughts on finding more datasets

Datasets appropriate for nonlinear regression modeling abound. If searching for more nonlinear datasets for which polynomial regression is appropriate, focus on physical relationships (e.g., the period of a pendulum or Boyle's Law), scientific formulae, etc. as the inspiration: many nonlinear relationships are well-established and appropriate for students learning regression. I chose to use fish data in this lesson because the overall three-dimensional size (volume) of fish is often approximated using a one-dimensional measure (length).[11]

Notes on Other Residual Plots

Students with more regression modeling experience may be expected to make residual diagnostic plots, such as those shown in the graph below. (This graph is called a 4-in-1 Plot and is produced by Minitab. It simply shows four residual diagnostic plots in a single graph. All four of these diagnostic plots were produced using the (regular) residuals, though students may use standardized residuals, too. The graphs in the 4-in-1 plot are: a normal probability plot of the residuals (top left), a histogram of the residuals (bottom left), a scatterplot of the residuals (Y-axis) versus the fitted values (X-axis), and a line plot with residuals on the Y-axis and the order of the observation in the dataset on the X-axis.)

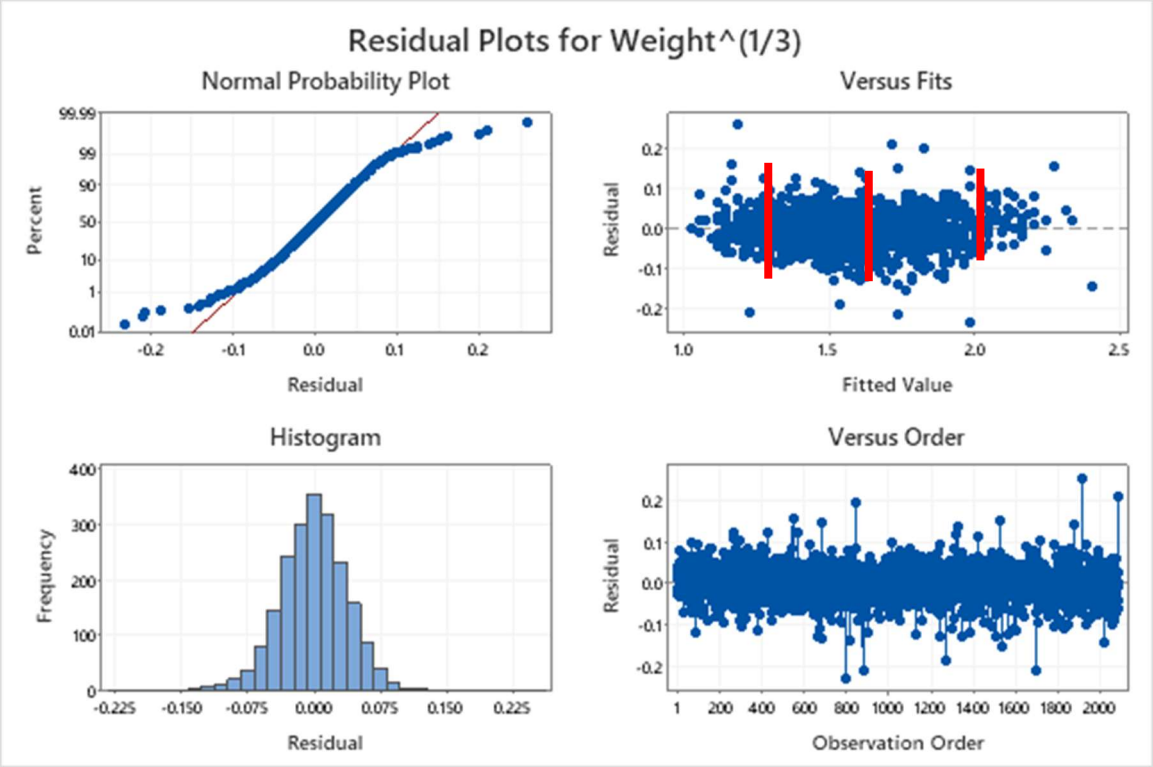
[11] Using fish as a context for nonlinear modeling has been done before in textbooks including *STAT2* (Cannon et al., 2013, 2018) and *The Practice of Statistics* (Starnes et al., 2012). The dataset used in this lesson has not previously been used.



The most important of these graphs for our initial modeling purpose is *Residuals vs. Fits*: a clear curved pattern is present, suggesting a nonlinear relationship between the variables. (Non-constant variance may also be an issue, though addressing the nonlinearity would be the first priority. To see the non-constant variance, note that the vertical distance between the largest and smallest residual at each point on the x-axis systematically changes, indicated by the red lines. Among the fish with the smallest length, there is less variability in weights; among fish that are of average length for the dataset there is considerable variability in weights. This observation about non-constant variance is a separate problem from the curvature.) The *Residuals vs. Fits* graph shows the same overestimation/underestimation mentioned before, but the addition of the horizontal dashed line at 0 might make this clearer.

For Model B (perhaps the most appropriate of the alternative models), the 4-in-1 Plot is shown below. Notice that both the curvature and non-constant variance apparent in the *Residuals vs. Fits* graph have been considerably improved relative to Model 0. In the *Residuals vs. Fits* graph – a plot of the residuals against the row number for each observation – there is an improvement in Model B relative to Model 0, too: in Model 0, there seems to be an asymmetry in the magnitude of the positive residuals when compared to the negative residuals (note that 0.0 is *not* in the middle of the Y-axis scale, and there is no need for -5.0 to be labeled on the Y-axis scale). However, with Model B this asymmetry appears to be improved: the magnitude of the positive residuals and negative residuals seem to be about the same (note that 0.0 *is* in the middle of the Y-axis scale and there seems to be roughly equal scattering above and below $y = 0$). Lastly, the distribution of the residuals for Model B more closely resembles the normal distribution. While not critical to the analysis in this lesson, the distribution of the residuals becomes important when inferential methods are used. For Model 0, the distribution of the residuals appears to be somewhat right skewed, but the distribution of the residuals for Model B appears to be approximately bell-shaped. While the distribution of the residuals does not appear to be exactly normal (as can be seen with the departures from the straight-line pattern in the *Normal Probability Plot*), the skew apparent in Model 0 is no longer present – yet another improvement. The *Versus Order* line plots for both Model 0 and Model B do not reveal potential patterns or relationships

between the residuals and the order the observations appeared in the dataset; if such a pattern existed, it would warrant further investigation. Note that the asymmetry in the magnitude of the residuals apparent in the *Residuals vs. Fits* graph for Model 0 and the improvement in this area for Model B can also be seen in the *Versus Order* graphs.



Examining the 4-in-1 Plots for the other models will reveal similar results: all six models represent an improvement relative to Model 0, but Models B, D, and F tend to represent greater improvement than Models A, C, and E.

Further Reading

This lesson touches on many regression topics that are emphasized (or covered) in introductory statistics courses: you may find it helpful to use a regression-focused textbook as a reference. While there are many such books available, I reference four here to point you in the right direction. Cannon et al. (2018) and Mendenhall and Sincich (2012) are regression books that focus on applications and have minimal prerequisites. Kutner et al. (2005) and Bingham and Fry (2010) include both applications and theory; these books routinely draw on knowledge of calculus and linear algebra. Every topic covered in this lesson is included in each of these books. While not as comprehensive as a textbook, another resource to consider are the course notes for Penn State's *STAT 501: Regression Methods* course; these notes are freely available online (see Lesson 9: Data Transformations[12]).

[12] The Pennsylvania State University, *STAT 501: Regression Methods*, Lesson 9: Data Transformations https://online.stat.psu.edu/stat501/lesson/9

References

Bingham, N. H., & Fry, J. M. (2010). *Regression: Linear models in statistics*. Springer.

Cannon, A. R., Cobb, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., Rossman, A. J., & Witmer, J. A. (2013). *STAT2: Building Models for a World of Data*. W. H. Freeman and Company.

Cannon, A. R., Cobb, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., Rossman, A. J., & Witmer, J. A. (2018). *STAT2: Modeling with Regression and ANOVA* (2nd ed.). W. H. Freeman and Company.

Environment and Climate Change Canada. (2019). *Wild Fish Health, Oil Sands Region*. http://data.ec.gc.ca/data/substances/monitor/fish-health-toxicology-contaminants-oil-sands-region/wild-fish-health-oil-sands-region/

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

Mendenhall, W., & Sincich, T. (2012). *A second course in statistics: Regression analysis* (7th ed.). Prentice Hall.

The Pennsylvania State University (n.d.). *STAT 501: Regression Methods* [online course]. OPEN.ED@PSU. https://online.stat.psu.edu/stat501/

Starnes, D. S., Yates, D. S., & Moore, D. S. (2012). *The practice of statistics* (4th ed.). W. H. Freeman.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913