OpenIntro online supplement

Fitting models for nonlinear trends

Prerequisites: Sections 1.1-1.6, 3.1, 4.1-4.4, 7.1-7.4, and 8.1 from OpenIntro Statistics are the bare minimum.

Figure 1 presents two examples of nonlinear relationships between two numerical variables. We'll introduce two techniques for fitting these two data sets: (1) transforming the response variable and (2) fitting nonlinear model using polynomial terms in multiple regression. While these two methods are very useful, there is no "one size fits all" modeling solution, and there are plenty of situations where these two methods will be insufficient for your needs. If you find that nonlinearity or challenges with residuals cannot be adequately addressed using these methods, consider turning to additional statistical methods.[1]

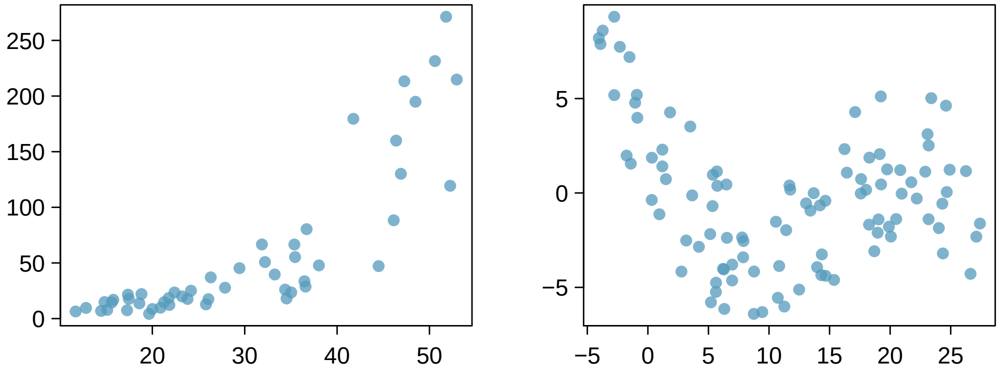

Figure 1: Two pairs of numerical variables where each relationship is nonlinear. The residuals may also show other deviations that must be considered when modeling these data, including non-normal or **heteroskedastic** residuals (*heteroskedastic* means *non-constant variance*).

The techniques introduced in this section may be useful when the first, second, or fourth conditions for a simpler linear model are violated:

1. the model residuals should be nearly normal,
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

1 Transformations on the response

Consider the scatterplot in the left panel of Figure 1. Here, the response y (vertical) tends to be positive but grow quickly. Additionally, the residuals show non-constant variance, because they are more variable for larger values of x (horizontal) and y. These two characteristics of the untransformed data are a clue that a transformation may be useful.

In Section 1.6 of OpenIntro Statistics, we learned about the power of transformations to make skewed data more symmetric. If we look at a histogram of the x and y variables in Figure 2, we can see that x shows a very slight right skew and y is strongly right skewed.

[1] See the Supplement page on openintro.org for recommended free books that may be useful, or post a question on the online Public Forums on openintro.org.

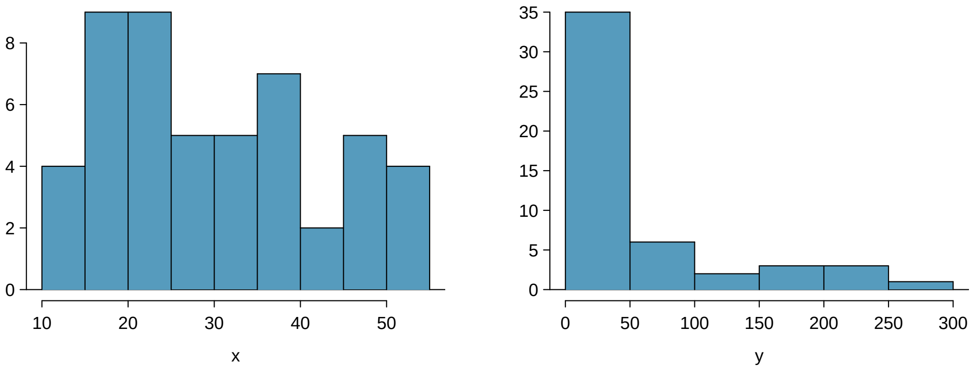

Figure 2: Histograms for both the x and y variables from the left panel of Figure 1.

This suggests that it may be useful to transform the y variable. Had x been strongly right skewed, then we should have also considered using a transformation on x.

There are many possible transformations, but one of the most common is the natural log-transformation (sometimes written as ln). We'll take the natural log for the y values and call this new variable $y^\star$:

$$y^\star = \log y, \text{ where "log" is the natural log}$$

Figure 3 shows $y^\star$ plotted against x. The data now show a linear relationship, where outliers are limited and the variability is roughly constant. Such an outcome is ideal, though far from guaranteed.

We may now readily fit a linear model to the transformed scatterplot:

$$\begin{aligned} \hat{y}^\star &= 1.03 + 0.08x \\ y^\star &= 1.03 + 0.08x + residuals \end{aligned}$$

In the first equation above, the formula has been written in the form used by OpenIntro Statistics. The second line is a more general way to write this formula. This general form is important when we are transforming data since we often want to **back-transform** the data. Here, we back-transform by substituting $\log(y)$ for $y^\star$ and then solve for y:

$$\begin{aligned} y^\star &= 1.03 + 0.08x + residuals \\ \log(y) &= 1.03 + 0.08x + residuals \\ y &= e^{1.03 + 0.08x + residuals} \end{aligned}$$

In this way, we can now enter a value for x and get an estimate for what value we think y will take. This fitted line is shown in Figure 4.

The predicted value for y in this model should *not* be confused with the expected (or mean) value of y for a given value of x, though the result may be somewhat close. The footnote provides an explanation of the difference for the interested reader.[2]

[2] Suppose we collected many observations where $x = 35$. This model suggests that the distribution of the corresponding y values would be skewed as a result of the relationship between the residuals and the outcome y being nonlinear. The model (roughly speaking) estimates the median for each value of x. Because the median is not the same as a mean in a skewed distribution, the model will not provide the expected value of y, though often times it will be close.

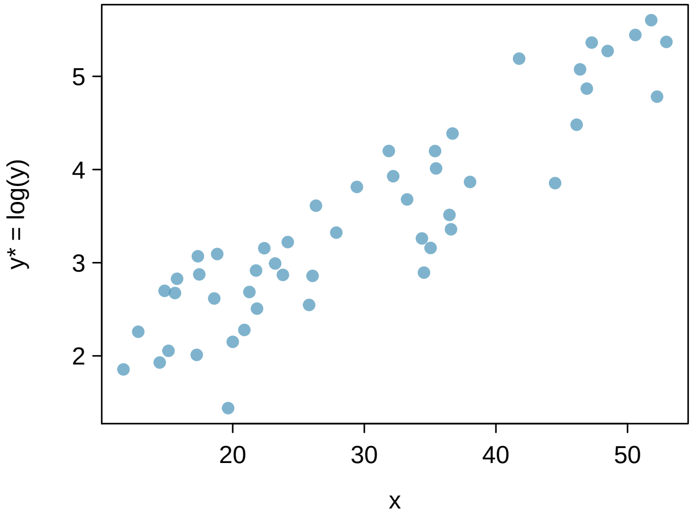

Figure 3: A plot of $y^\star$ (the result of transforming y by taking the natural logarithm $\log y$) against x. The relationship between $y^\star$ and x appears to be linear.

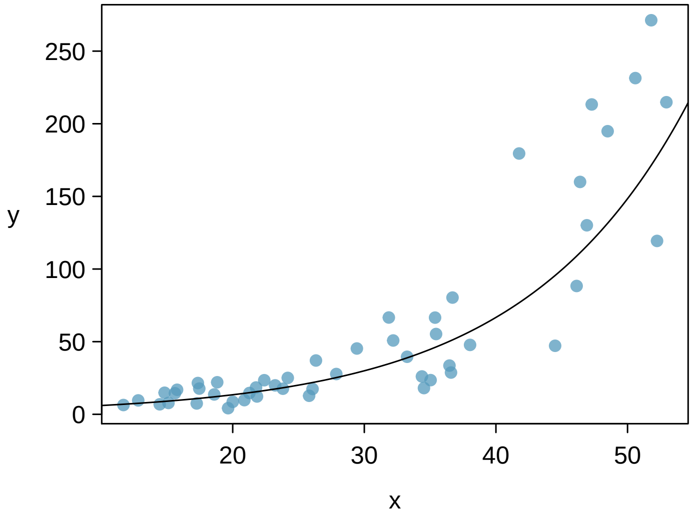

Figure 4: A nonlinear curve through the data generated by fitting a model of the form $\log y = \beta_0 + \beta_1 x + residuals$, then solving for y.

TIP: Interpreting coefficients from a model that used $\log y$
If the outcome in a model was transformed using the natural logarithm and the model fits well, then y tends to grow (or decay) **exponentially** relative to x.

Caution: Transformations can be abused
There is a very large number of possible transformations. If we keep trying transformations until one "works", we have not effectively modeled our data. Rather, we have performed a complicated form of data fishing where we mine the data until we see structure. This apparent structure may just be due to chance. Therefore, think carefully about transformations before applying them.

You are once again armed with knowledge that is both powerful and dangerous. This very brief introduction to transformations should be useful for informal projects. For a more complete review of this topic, visit Chapter 8 of *Practical Regression and ANOVA in R*, which can be found in the Free Books section on the Supplements page of openintro.org.

2 Fitting a polynomial curve

Let's take a look at the second nonlinear relationship we saw in Figure 1, which appears again in Figure 5 with a poorly-fit straight line. Here we see what appears to be a nonlinear relationship but where the residuals would be approximately **homoskedastic** (constant variance) if we could reasonably model the curve of the line. This is a good signal that we want to fit a curve but not perform a transformation. We can do so by generating a **polynomial basis** of x: $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, and so on. In short, we will use the variables x_1, x_2, x_3, ... in a multiple regression model instead of simply the original variable x. We should note that it is uncommon to use terms beyond $x_2 = x^2$ and very rarely beyond $x_3 = x^3$.

We start by fitting a linear model to the data, where the best-fitting straight line is shown in Figure 5 and summarized as

$$y = 0.8441 - 0.0964x + residuals$$

Even without checking the residual plot, it is evident that this line does not fit the data well, though Table 6 shows that the linear term is statistically significant.

⊙ **Exercise 1** Suppose you were providing feedback to someone on a project, and the colleague had fit the line to the data shown in Figure 5. Suppose also that your colleague believes this model is sufficient because the estimate for the slope is statistically significant. Explain why the model is inappropriate. One possible explanation is provided in the footnote.[3]

As a next step, we'll add another variable to the model: $x_2 = x^2$. This new variable is itself a transformation on the variable x. However, rather than substituting x_2 for x, we'll fit a multiple regression model *including both variables* from the polynomial basis:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + residuals$$

[3] Regression models require certain conditions to be met. In particular, the residuals must be independent of each other. However, when we look at the residuals from the fit in Figure 5, there is a clear trend in the residuals not captured by the straight line, meaning the independence condition is violated and the model is inadequate.

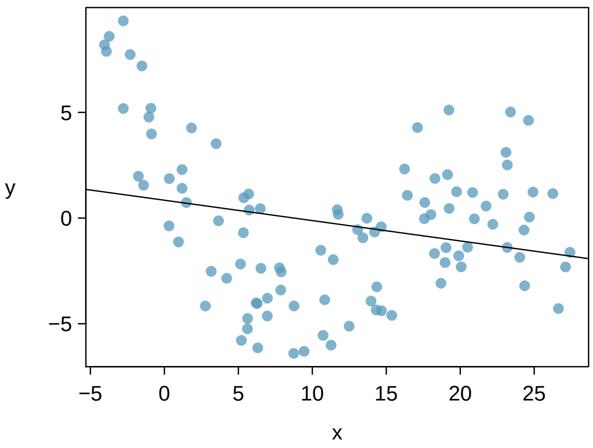

Figure 5: Scatterplot with the best-fitting straight line, which does not fit the data well.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8441	0.5799	1.46	0.1487
x1	-0.0964	0.0397	-2.43	0.0169

Table 6: Summary for a straight line fit to the data shown in Figure 5.

The best-fitting model of this form is shown in Figure 7, and the summary for the model is shown in Table 8.

● **Example 2** Write out the best fitting quadratic model using Table 8.

The model may be written as

$$\begin{aligned} y &= 2.4252 - 0.7769x_1 + 0.0295x_2 + residuals \\ &= 2.4252 - 0.7769x + 0.0295x^2 + residuals \end{aligned}$$

In this example a quadratic model is still insufficient, so we will try a cubic polynomial (also known as a third-order polynomial). We will try fitting a model based on a cubic polynomial:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + residuals \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + residuals \end{aligned}$$

Such a model is summarized in Figure 9 and Table 10.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.4252	0.5079	4.78	0.0000
x1	-0.7769	0.0956	-8.13	0.0000
x2	0.0295	0.0039	7.55	0.0000

Table 8: Summary for a quadratic fit to the data shown in Figure 7.

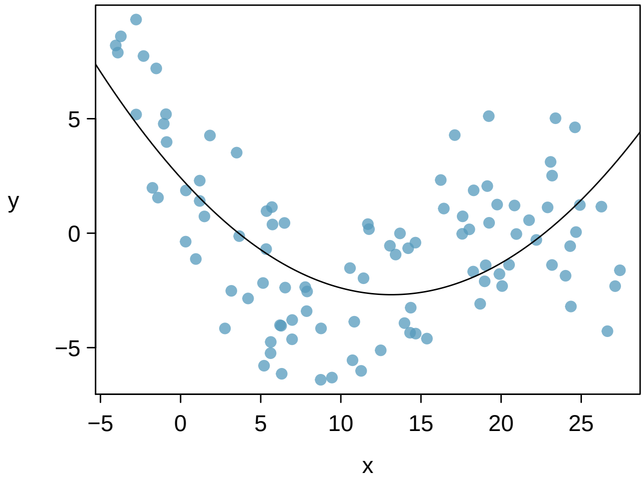

Figure 7: Scatterplot with the best-fitting quadratic line, which fits better than a straight-line but still misses some data structures. For example, the model underestimates much of the data for the range -5 to 0 and it overestimates nearly all of the data between 5 and 10.

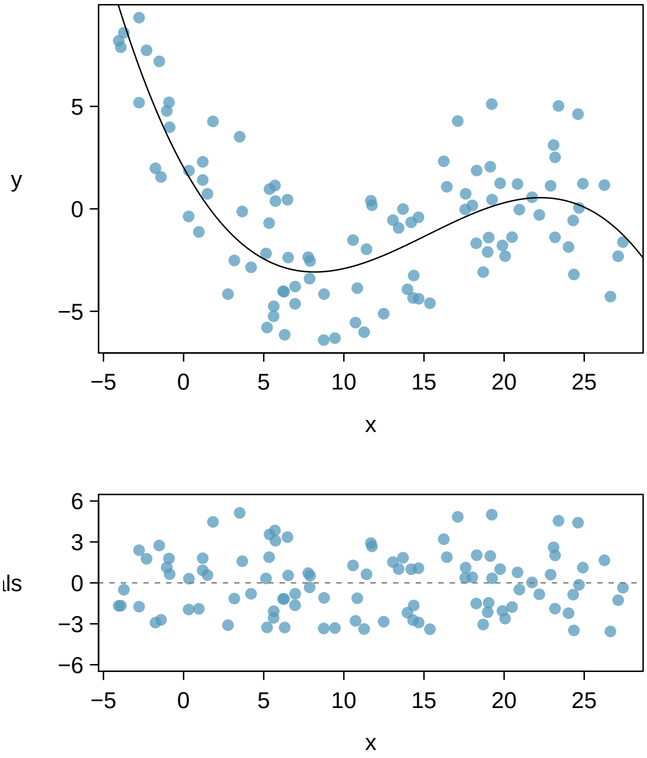

Figure 9: Scatterplot with the best-fitting cubic line. The residual plot shows no apparent structure, which is a good sign the model is fitting well.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0187	0.4242	4.76	0.0000
x1	-1.4202	0.1236	-11.49	0.0000
x2	0.1187	0.0136	8.75	0.0000
x3	-0.0026	0.0004	-6.77	0.0000

Table 10: Summary for a cubic line fit to the data shown in Figure 9

Exercise 3 Write out the best fitting quadratic model using Table 10. The solution is in the footnote.[4]

The initial prognosis from the residual plot is that the cubic model fits very well. However, a complete analysis would include checking the model diagnostics carefully, which is a topic discussed in Section 8.3 of OpenIntro Statistics.

TIP: Stick with lower-order polynomials

If you want to try out using a polynomial term in your model, consider x^2 and perhaps x^3 if the model is still not a good fit. If a cubic polynomial will not model your data well, then be very cautious about trying higher-order polynomials. Instead, consider learning about regression splines, kernel smoothing, or some other statistical technique. See the textbook Elements of Statistical Learning for more information on advanced modeling techniques.

Caution: Do not extrapolate with transformed models or models that use polynomial terms

Extrapolation is already treacherous for any model, but it can be much worse for transformed data or data that includes polynomial terms, as the model can deviate very rapidly from the typical values observed in the original data set.

[4] $y = 2.0187 - 1.4202x_1 + 0.1187x_2 - 0.0026x_3 + residuals = 2.0187 - 1.4202x + 0.1187x^2 - 0.0026x^3 + residuals$