Section 4.2 Fitting Curves and Surfaces by Least Squares

4.2.1 Curve Fitting

In many cases the relationship of $y$ to $x$ is not a straight line.

To fit a curve to the data on can

1. Fit a nonlinear function directly to the data.
2. Rescale, transform $x$ or $y$ to make the relationship linear.
3. Fit a polynomial function to the data.

For a uniform fluid
- light should decay as an exponential function of depth
- Given data on depth, $x$, and light intensity, $y$,
  - the data should look like an exponential function.
  - $y \approx ae^{bx}$

1. Direct Fitting

We could fit the equation directly by

- minimizing $\sum \left( y_i - ae^{bx_i} \right)^2$
- If the residuals are normal with equal variance then
  - This is a fine way to go.
- But more often the variances are not constant.
  - This is not the best approach.
A squared error of $(\pm 5)^2$ between the fitted and measured values should count as the same discrepancy for $y=20$ where variances are large as for $y=3$ where variances are small.

To do the direct nonlinear fitting
- We could use a "nonlinear regression" program to accomplish this.
- This can also be accomplish with the "Solver" add-in in Excel
- We won't be doing such curves this way in Stat 3411.

But very often in situations such as exponentially curved data
- the variances are not constant.
- One way to handle nonconstant variance is "weighted regression".
  - We give larger weight to residuals
    - where variance is small
    - we are more sure of where the fitted line should go.

We won't be doing this in Stat 3411.

2. Transforming

Often with an exponential relationship
- the larger values have more variance and
- the smaller values have smaller variances.

In this case it can work better to rescale the data

\[ y \approx ae^{b_1x} \]
\[ \ln(y) = \ln(a) + b_1x_1 \]
\[ y^{new} = y' = b_0 + b_1x_1 \]
Often values in the log scale have fairly constant variances.

A residual plot gives us a magnified view of the increasing variance and curvature.

This residual plot indicates 2 problems with this linear least squares fit
- The relationship is not linear
  - Indicated by the curvature in the residual plot
- The variance is not constant
  - So least squares isn't the best approach even if we handle the nonlinearity.

Don't fit an exponential function to these data directly with least squares.
- With increasing variability, not all squared errors should count equally.
  - A squared error of $5^2$ for Predicted=1
  - Is a bigger problem than a squared error of $5^2$ for Predicted=16.
Fit a line to these data.
- Here is makes sense for all residuals to count equally.
Then back-transform to the original scale.

**Extrapolation vs Interpolation**

Predicted values are
- less reliable
- if you are extrapolating outside the range of x values in the data set.

Extrapolations are more reliable
- if one has an equation from a physical model.
- For example physical models predict that in a uniform fluid light will decay exponentially with depth.
3. Polynomials

When we have no theory to guide us,
- We can often fit the curve in the range of observed x values
- With a polynomial function.

For example a cubic polynomial would be

\[ y \approx b_0 + b_1 x + b_2 x^2 + b_3 x^3 \]

This is linear function for the three variables

\[ x_1 = x \quad x_1 = x^2 \quad x_3 = x^3 \]

\[ y \approx b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \]

Excel and other programs fit these sorts of polynomial models.

In Excel
- Add new columns for \( x^2 \) and \( x^3 \)
- Run a regression using 3 columns for the x variables
  - \( x, x^2, \) and \( x^3 \)
  - All three columns need to be next to each other in Excel.
Given the fitted function, we want to check for an adequate fit

- by plotting the data
- along with the fitted function.

Notice that replication is useful

- for assessing how well the functional form fits the data.
- With replication here we can tell that the quadratic polynomial is underfitting the y values at x = 2.

We would plot residuals in the same ways suggested in section 4.1

- Versus predicted values $\hat{y}$
- Versus x
- In run order
- Versus other potentially influential variables, e.g. technician
- Normal Plot of residuals
Section 4.2.2 Surface Fitting by Least Squares

In many situations the response variable, y, is
- affected by more than one x variable.

For example we could have (see problem 21 in the Exercises)

\[
y = \text{armor strength} \\
x_1 = \text{thickness} \\
x_2 = \text{Brinell hardness}
\]

The simplest model to fit in this case is a linear model in both variables

\[
y \approx b_0 + b_1 x_1 + b_2 x_2
\]

In this case we are fitting a plane to the 3-D points.

This sort of linear model with
- more than one x variable is called
- "multiple linear regression"
For this function fitted y values
- for a fixed value of $x_1$
- Follow parallel lines when plotted against $x_2$.

\[
y_i \approx b_0 + b_1 x_{1i} + b_2 x_{2i}
\]

Again, we find $b_0$, $b_1$, $b_2$ to minimize
- the sum of squared deviations of the fitted values,
- the residual or error sum of squares.

\[
\min \sum (y_i - b_0 + b_1 x_{1i} + b_2 x_{2i})^2
\]

This minimization can be solved with explicit matrix formulas. Many programs including Excel have the capability.
There can be any number of x variables, for example

\[ y \approx b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \]

It is also possible to introduce curvature into one or more of the variables

For example \( y \approx b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_2^2 + b_5 x_3^2 + b_6 x_2 x_3 \)

Product terms (called interactions) in the model give nonparallel contours.
Residuals

Residuals are as always \( e_i = y_i - \hat{y} \).

\( R^2 \) is still the percent of variation explained by the regression.
- How better do we do
  - using the regression equation versus
  - using the mean to predict all values

\[
R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \times 100
\]

\( R^2 \) is also the square of the correlation between \( y_i \) and \( \hat{y}_i \).

To check how well the model fits we
- plot residuals in the same ways as earlier
  - Versus predicted values \( \hat{y} \)
  - Versus x
  - In run order
  - Versus other potentially influential variables, e.g. technician
  - Normal Plot of residuals
- In addition we would plot the residuals
  - Versus each x variable
    - Versus \( x_1 \)
    - Versus \( x_2 \)
    - Versus \( x_3 \)
    - An so on
Final Choice of Model

We have many choices for our final model including

- Transformations of many kinds
- Polynomials

Our final choice is guided by 3 considerations

1. Sensible engineering models and experience
2. How well the model fits the data.
3. The simplicity of the model.

1. **Sensible engineering models and experience**
   - Whenever possible use models that make intuitive physical sense
   - Good mathematical models based on physics makes extrapolation much less dangerous.
   - Having physically interpretable estimated parameters

2. **Obviously, the model needs to fit the data reasonably well.**
   - **Always** plot data and residuals in informative ways.

3. **Simplicity of the model.**

A very complex model relative to the amount of data

- Can fit the data well
- But not predict new points well, particularly extrapolated points.
- "Overfitting" the model
- Learn more in Stat 5511 Regression Analysis
The fourth order polynomial in pink
- fits the data exactly,
- but likely would not work well for predicting for $x=0.5$ or $x=1.5$

The linear fit in blue
- likely predicts new points better.

Often transforming to a log scale allows simpler models to fit the data reasonably well.
Extrapolating

With multiple x variables we need to be careful about
- extrapolating beyond the region of x variables for the observed data.
- We can't necessarily tell that a combination of X1 and X2 is unusual
- just by seeing
  - separately
  - where the new value of X1 falls amongst the measured X1 values
  - where the new value of X2 falls amongst the measured X2 values
Some Study Questions:

You run a regression model fitting $y$ with $x_1$, $x_2$, and $x_3$. What five residual plots would you need to check in order see if the data are fit well by the model?

What is extrapolation? Why is extrapolation potentially risky?

What can we do to reduce potential risks in extrapolating. (See pages 13 and 14.)

We run a regression predicting $y =$ armor strength from $x_1 =$ thickness and $x_2 =$ Brinell hardness and obtain $R^2 = 80\%$. Explain in words for a manager who has not seen $R^2$ values what it means for $R^2$ to be 80\%.

What is the problem with fitting a complex model to a small data set in order to fit the data points very exactly?
The following residual plot is for predicting product lifetimes from using $x_1$ and $x_2$.

What 3 problems with our fitting method and results are indicated by this residual plot?
What might you do to try finding a better fit to the data?