Math/Stat 2300 **Linear Regression (6.3)** (April 6)

from text *A First Course in Mathematical Modeling*, Giordano, Fox, Horton, Weir, 2009.

Linear regression continues our discussion of Least Squares criteria for fitting models to collected data.

Least Squares criteria minimizes the sum of the squared deviations, but is only for a single observation y_i for each value of the independent variable x_i. For multiple observations, we use linear regression.

We are only going to consider the basic concepts and interpretations. More indepth studies are left for a mode advanced statistics course. Our objectives are

- illustrate basic linear regression model
- define and interpret R^2
- interpret the residual plots

The Linear Regression Model

The basic linear regression model is

$$y_i = ax_i + b, \quad i = 1, \ldots, m$$

where there are m data points.
Previously, we have derived the normal equations

$$\begin{aligned} a\sum_{i=1}^{m} x_i^2 + b\sum_{i=1}^{m} x_i &= \sum_{i=1}^{m} x_i y_i \\ a\sum_{i=1}^{m} x_i + mb &= \sum_{i=1}^{m} y_i \end{aligned}$$

and we solved these to obtain

$$\begin{aligned} \text{slope } a &= \frac{m\sum x_i y_i - \sum x_i \sum y_i}{m\sum x_i^2 - \left(\sum x_i\right)^2} \\ \text{intercept } b &= \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{m\sum x_i^2 - \left(\sum x_i\right)^2} \end{aligned}$$

Now, we add other equations to help in our analysis.

- **Sum-of-squares error** is given by

  $$\mathrm{SSE} = \sum_{i=1}^{m} [y_i - (ax_i + b)]^2$$

  It reflects variation about the regression line.

  The smaller SSE, the better the approximating function fits the data.

- **Total corrected sum of squares** of y given by

  $$\mathrm{SST} = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

  where $\bar{y}$ is the average of the y values for the data points (x_i, y_i), $i = 1, \dots, m$

  Note that $\bar{y}$ is also the average value of the linear regression line $y = ax + b$ over the range of data.

- **Regression sum-of-squares** given by

  $$\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}$$

  SSR reflects the amount of variation in the y values explained by the linear regression line $y = ax + b$ when compared with the variation in the y values about the line $y = \bar{y}$.

Note that $\mathrm{SST} \geq \mathrm{SSE}$.

We define the coefficient of determination as R^2, which is a measure of fit for the regression line,

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

The number R^2 expresses the proportion of the total variation in the y variable of the actual data (when compared to the line $y = \bar{y}$) that can be accounted for by the straight-line model, whose values are given by $ax + b$ (the predicted values) and calculated in terms of the x variable.

Note: the closer the value of R^2 is to 1, the better the fit of the regression line model to the actual data.

Some properties of the statistic R^2:

- $R^2 \leq 1$
- the value of R^2 does not depend on which of the two variables is labeled x and which is labeled y
- the value of R^2 is independent of the unit of x and y

Residuals

Another way to evaluate the reasonableness of the fit of the model is to look at the plot of residuals versus the independent variable.

Recall: the **residuals** are the errors between the actual and predicted values

$$r_i = y_i - f(x_i) = y_i - (ax_i + b)$$

If we consider the plot of the residuals vs the independent variable, we obtain some valuable information:

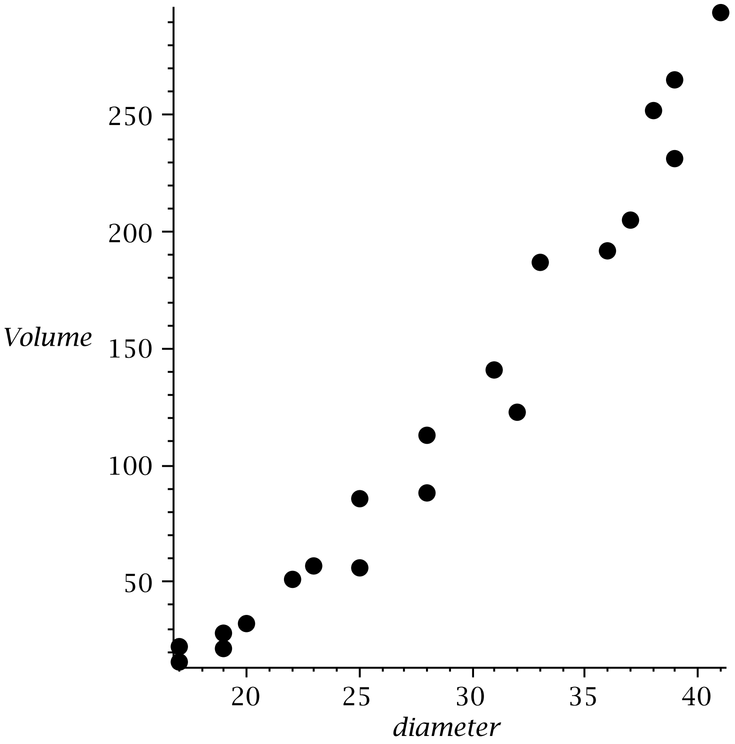
- The residuals should be randomly distributed and contained in a reasonable small band that corresponds with the accuracy of the data.
- An extremely large residual should indicate that we need to take a further look at the associated data point to discover the cause of the large residual.
- A pattern or trend in the residuals indicates that some effect or aspect on the data has not been captured in the model. The nature of the pattern can provide hints as to how to refine our model.

Example

Consider our ponderosa pine question from before. We had the following data:

Diameter	Board feet
36	192
28	113
28	88
41	294
19	28
32	123
22	51
38	252
25	56
17	16
31	141
20	32
25	86
19	21
39	231
33	187
17	22
37	205
23	57
39	265

Looking at the scatterplot of the raw data, the plot is concave up and increasing. So the trend in the data suggests a power function.



Recall that we had determined two possible models, by using geometric similarity,

$$V \propto d^3, \quad V \propto d^2$$

In addition, we could make some assumption about the volume associated with the underground root system: we assume that this volume is constant

$$V = ad^3 + b, \;\; V = \alpha d^2 + \beta$$

By fitting our data to those four different models, we determine the constants for each model:

$$\begin{aligned} V &= 0.00431d^3 \\ V &= 0.00426d^3 + 2.08 \\ V &= 0.152d^2 \\ V &= 0.194d^2 - 45.7 \end{aligned}$$

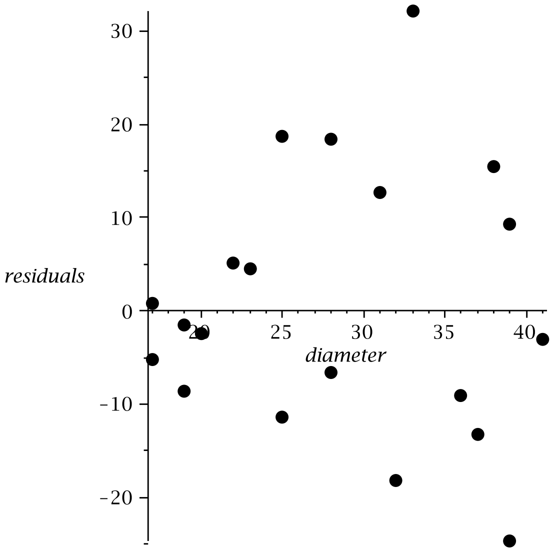
Applying linear regression, we obtain:

Model	SSE	SSR	SST	R^2
$V = 0.00431d^2$	3742	458 536	462 278	0.9919
$V = 0.00426d^3 + 2.08$	3712	155986	159 698	0.977
$V = 0.152d^2$	12 895	449 383	462 278	0.9721
$V = 0.194d^2 - 45.7$	3910	155 788	159 698	0.976

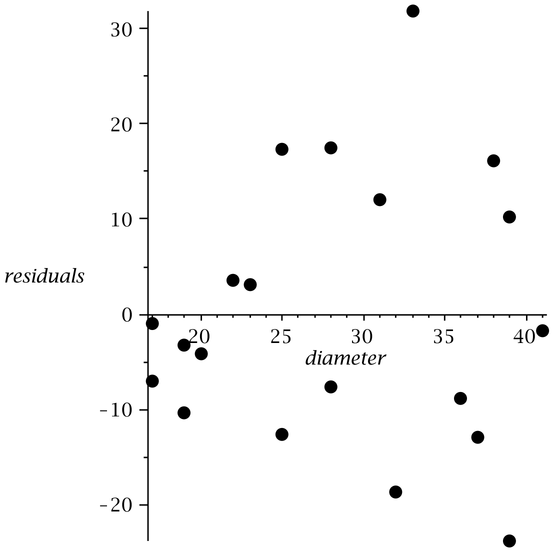
What do you think these statistics tell us?

Now, we plot the residuals for each of these models (the plots are on the next page).

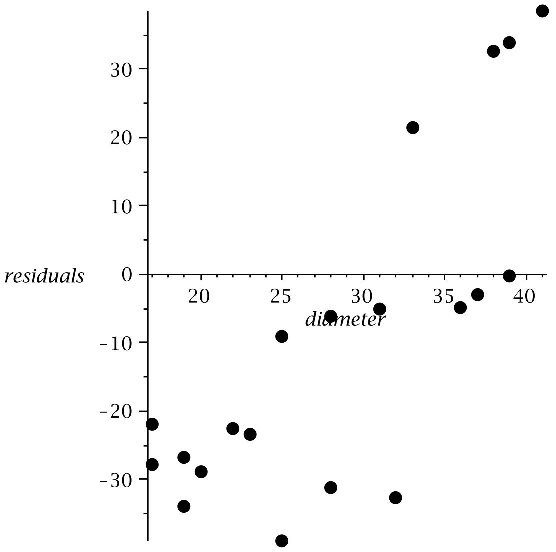
What do we notice about the residuals plot? What models are reasonable?



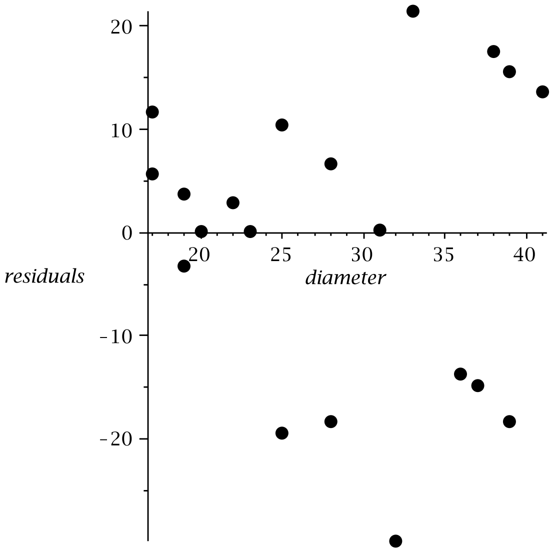
$V = 0.00431d^3$



$V = 0.00426d^3 + 2.08$



$V = 0.152d^2$



$V = 0.194d^2 - 45.7$