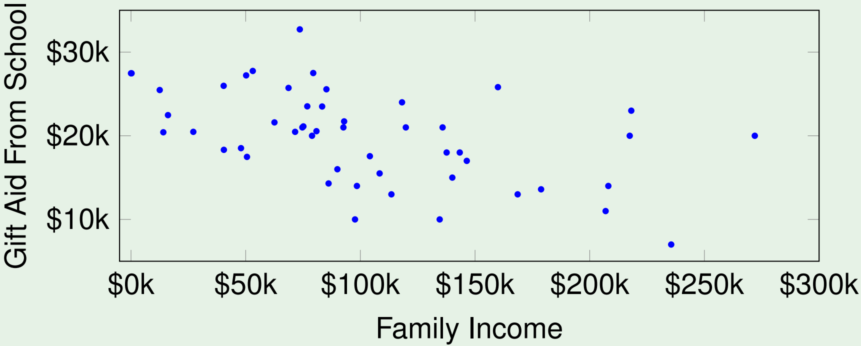
Least Squares Regression

Colby Community College

Example 1

Gift aid is financial aid that does not need to be paid back.

A sample of 50 random freshmen at Elmhurst College is shown, comparing the student's family income against gift aid received.



Which of the lines best fits the data?

Without an objective definition of measure of "best", the answer will vary from person to person.

What does "best" mean?

A reasonable idea of best, is if we make the sum of the residuals as small as possible:

$$|e_1| + |e_2| + \cdots + |e_n|$$

A more common practice is to choose a line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Definition

The line that minimizes the sum of the squares of the residuals is called the **least squares line**.

Note

In many applications, a residual twice as large as another is more than twice as bad. Squaring the residuals helps account for this discrepancy.

Conditions for the Least Squares Line

Linearity: The data should show a linear trend. If there is a non-linear trend a more advanced method is needed.

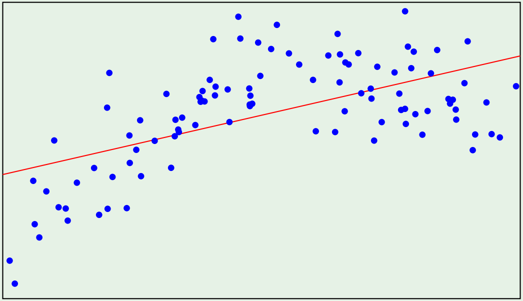
Near Normal Residuals: When this condition is found unreasonable, it is usually because of outliers or concerns about influential points.

Constant Variability: The variability of points around the least squares line remains roughly constant.

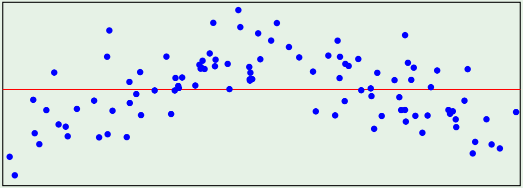
Independent Observations: Be careful about applying regression to **time series** data, which are sequential observations in time such as a stock price each day.

Example 2

Scatter plot where the linearity condition fails:

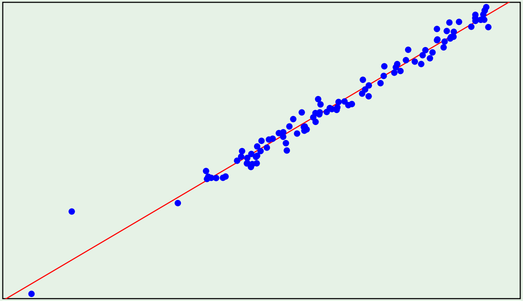


Residual plot:

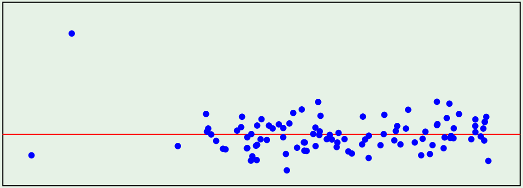


Example 3

Scatter plot where there are clear outliers:

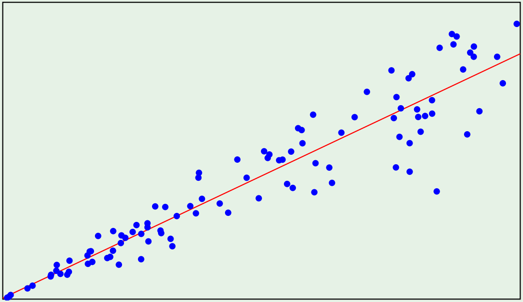


Residual plot:

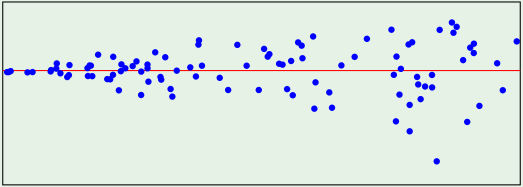


Example 4

Scatter plot where the variability around the line isn't constant:

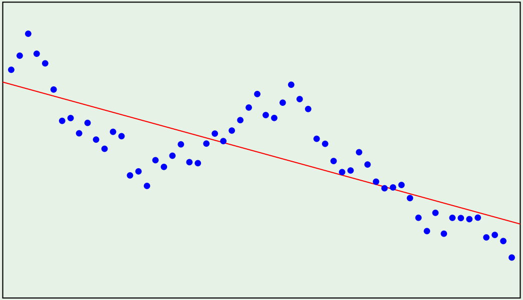


Residual plot:

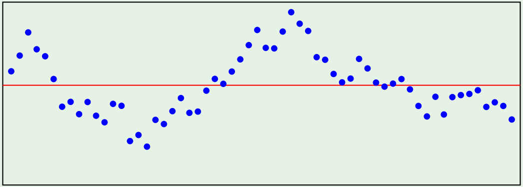


Example 5

Scatter plot using time series data:



Residual plot:



Finding the Least Squares Line

The least squares regression line will have form:

$$\hat{y} = \beta_0 + \beta_1 x$$

While technology is usually used to find β_0 and β_1, we can use the following properties to estimate them by hand:

- The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} R$$

- The point $(\bar{x}, \bar{y})$ is on the least squares line.

Note

Recall from Algebra that if we know the slope, m, of a line and a point, (x_0, y_0), on that line, then:

$$y - y_0 = m(x - x_0)$$

Example 6

The summary statistics of the Elmhurst College data set are:

	Family Income (x)		Gift Aid (y)	
mean	$\bar{x} =$	$\$101,780$	$\bar{y} =$	$\$19,940$
std. dev.	$s_x =$	$\$63,200$	$s_y =$	$\$5,460$

The correlation of the data set is $R = -0.499$, so

$$b_1 = \frac{s_y}{s_x} R = \frac{5,460}{63,200}(-0.499) = -0.0431$$

Since $(\bar{x}, \bar{y}) = (101,780, 19,940)$ is on the least squares line, we have $x_0 = 101,780$ and $y_0 = 19,940$ which gives

$$\begin{aligned} y - y_0 &= m(x - x_0) \\ y - 19,940 &= -0.0431(x - 101,780) \\ y - 19,940 &= -0.0431x + 4386.72 \\ y &= -0.0431x + 4386.72 + 19,940 \\ y &= 24,327 - 0.0431x \end{aligned}$$

Process for estimating the least squares line

1. Estimate the slope parameter: $b_1 = \frac{s_y}{s_x} R$
2. Since $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$
3. Using the point-slope form: $b_0 = \bar{y} - b_1 x$

Note

The slope, b_1, describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger.

Note

The intercept, b_0, describes the average outcome of y if $x = 0$ and the linear model is valid all the way to $x = 0$.

Example 6 (Continued)

The slope, $b_1 = -0.0431$, means that for each \$1,000 family income, we would expect a student to receive a net difference of

$$\$1,000 \cdot (-0.0431) = -\$43.10$$

Which means, on average, \$43.10 less in gift aid.

The intercept, $b_0 = \$24,319$, gives the gift aid, on average, a student would receive if their family had no income.

Note

We must be cautious about interpreting a causal connection between these variables because this data is observational, not experimental.

Definition

When a regression is used to predict from a x value in between the maximum and minimum values, it is called **interpolation.**

Definition

When a regression is used to predict from a x value bigger than the maximum or smaller than the minimum, it is called **extrapolation.**

Example 7

The largest family income in the Elmhurst data set is \$271,974.

If we use the least squares line to estimate the aid of a student with a family income of \$1,000,000, we would get:

$$24,319 - 0.0431(1,000,000) = -18,781$$

The financial aid a school gives a student can never be less than zero!

Strength of Fit

We have used the correlation R to describe the linear relationship between two variable, but it is more common to use R^2, called **R-squared**.

The R^2 of a linear model describes what percent of the variation in the response that is explained by the least squares line.

Example 8

With the Elmhurst College data, the variance of the response variable is $s^2_{\text{aid}} = (5,460)^2 \approx 29.8$ million.

If we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's family income.

The variability in the residuals describes how much variation remains after using the model: $s^2_{residuals} \approx 22.4$ million.

This means we have a reduction of

$$\frac{s^2_{aid} - s^2_{residuals}}{s^2_{aid}} = \frac{29,800,000 - 22,400,000}{29,800,000} = \frac{7,500,000}{29,800,000} \approx 0.25$$

Which means a reduction of about 25% by using information about family income for predicting aid.

Note that for this data set we have $R = -0.499$ and

$$R^2 = (-0.4999)^2 \approx 0.25$$