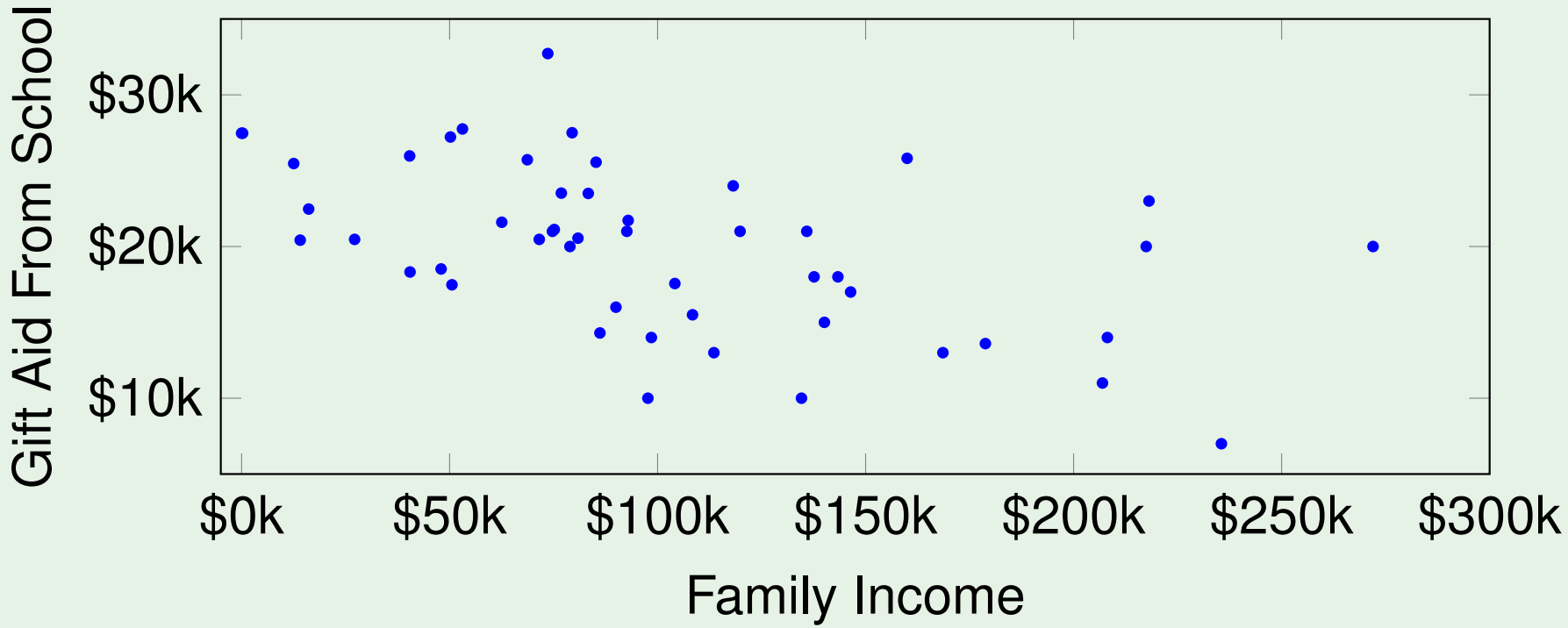
Least Squares Regression

Colby Community College

Example 1

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

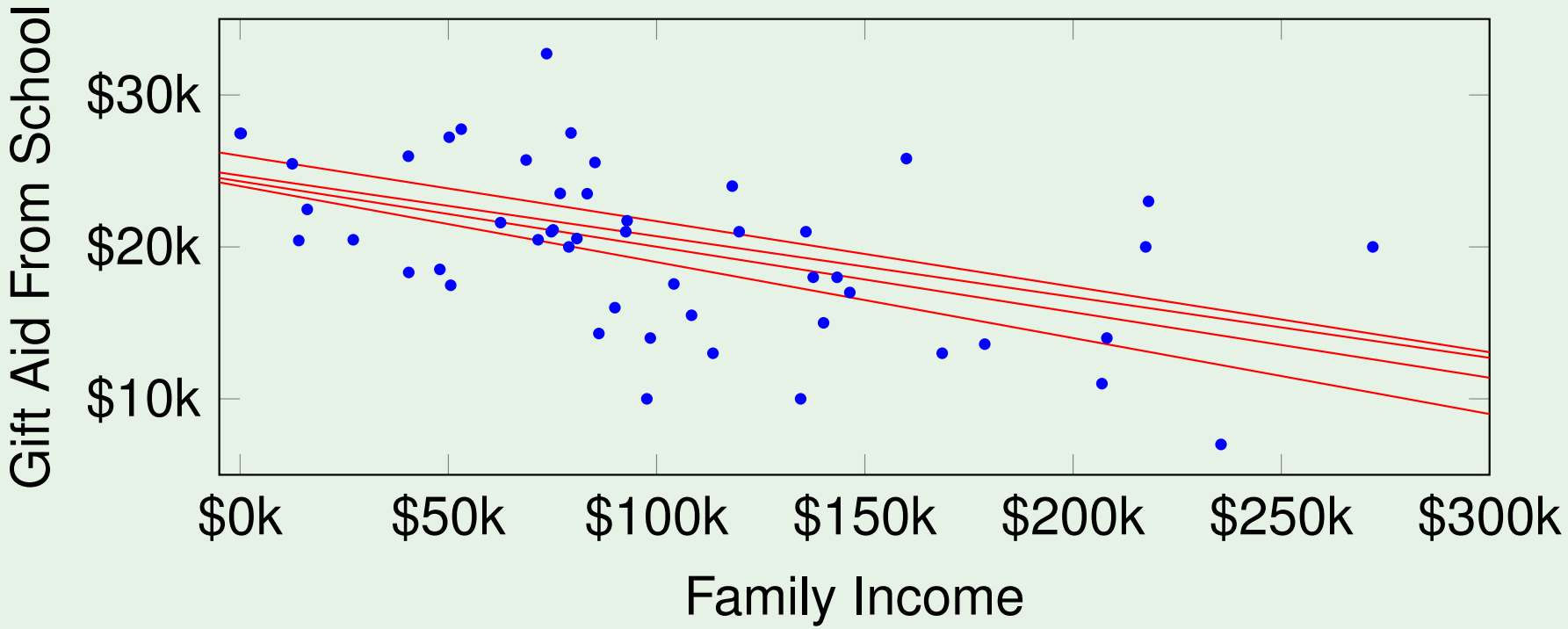
A sample of 50 random freshmen at Elmhurst College is shown, comparing the student's family income against gift aid received.



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.

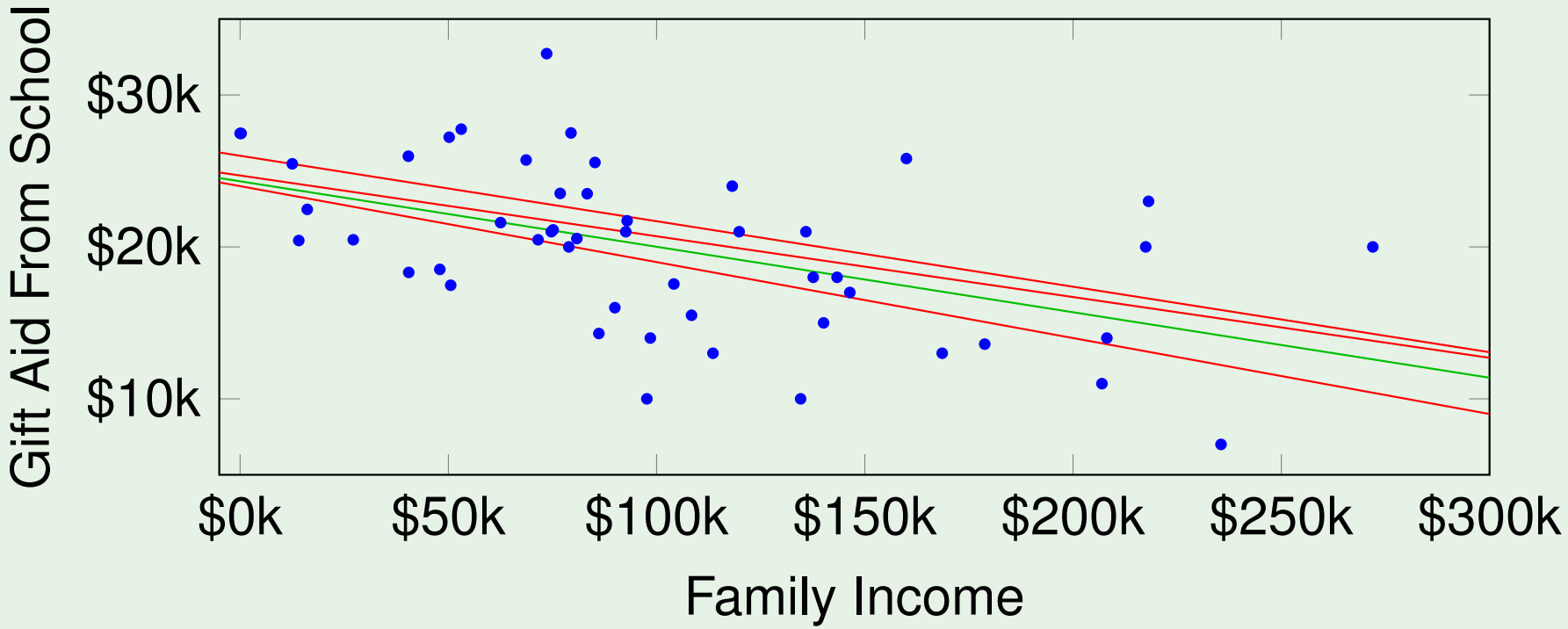


Which of the lines best fits the data?

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:

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

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:

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

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**.

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.

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.

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.

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.

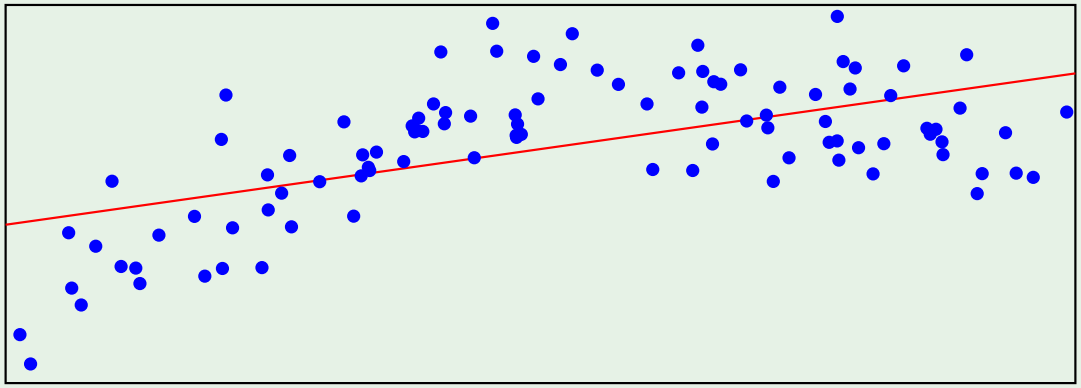
Near Normal Residuals: When this condition is found unreasonable, it is usually because of outliers or concerns about influential points.

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 linearity fails:



Residual plot: