Probability and Mathematical Statistics in Data Science

Lecture 34: Section 12.1: The Simple Linear Regression Model cont.

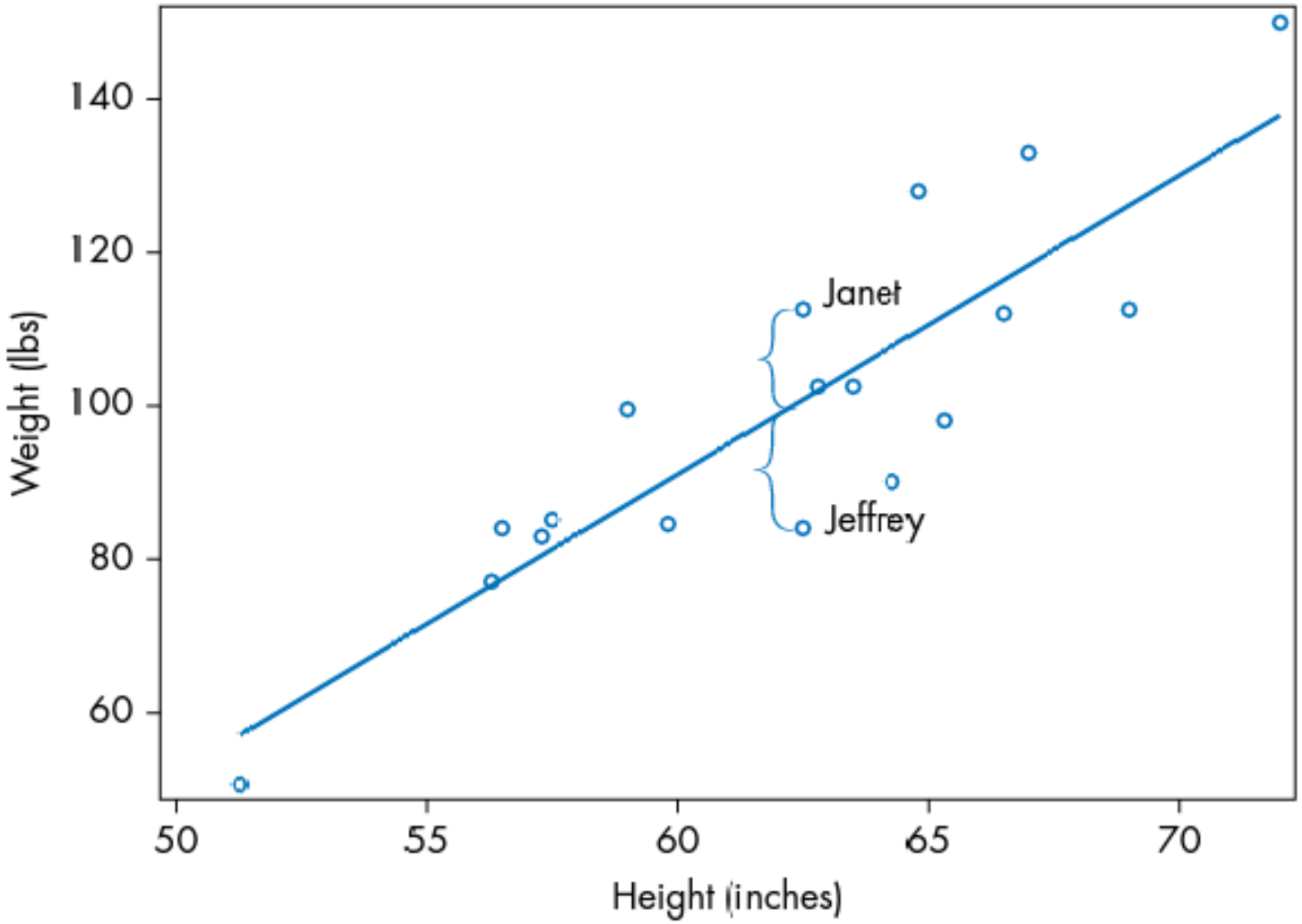
Recap: Deviations around the Line of Best Fit

- To understand how the line of best fit (regression line) is chosen, we need to understand residuals.

- A residual is a measure of the vertical distance of an individual's value of the response variable to the predicted value of the response variable

- The regression line can be thought of as a line of means. Therefore, we can think of a residual as the deviation of a individual value from the mean (predicted value)

Example: 19 Children: Height and Weight



19 Children Height-Weight Example

There are two children in our sample, Janet and Jeffrey, with a height of 62.5 inches. The individual observed weights for Janet and Jeffrey are 112.5 lbs. and 84 lbs., respectively.

$$\text{Predicted Weight} = -143 + 3.9 \times (62.5)$$
$$= -143 + 243.75$$
$$= 100.75 \text{ lbs.}$$

Therefore, the individual residual deviations for Janet and Jeffrey are as follows:

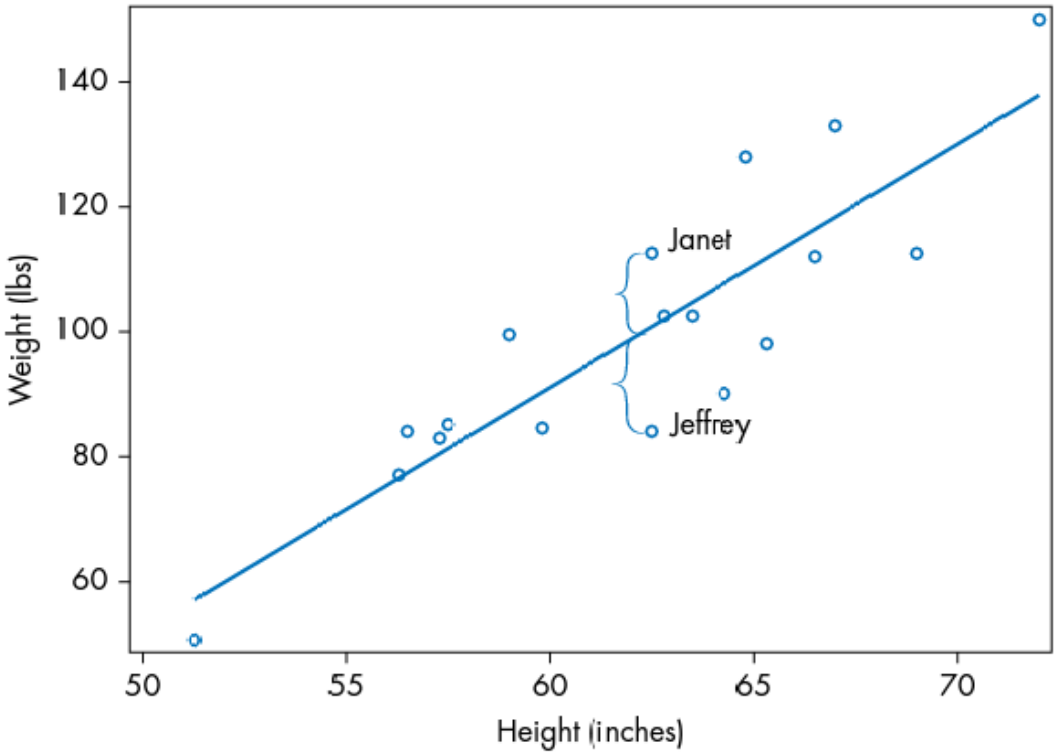
$$e_i = y_i - \hat{y}_i$$

residual deviation = observed weight – predicted weight

Janet: 112.5 lbs.: residual deviation = 112.5 – 100.75 = 11.75 lbs.

Jeffrey: 84 lbs.: residual deviation = 84 – 100.75 = –16.75 lbs.

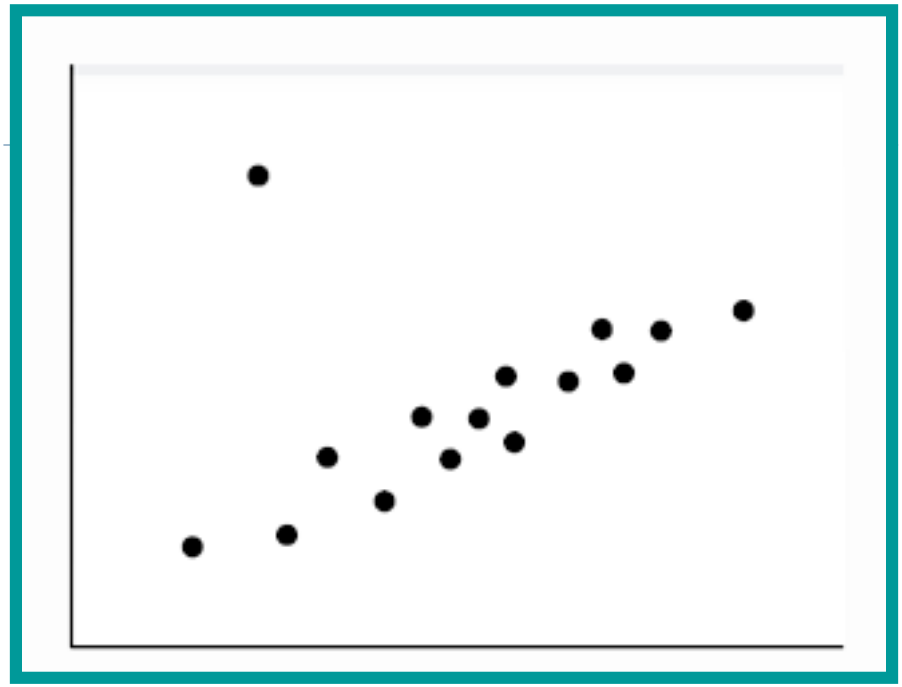
19 Children Height-Weight Example



$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Line of Best Fit -> Minimize the Sum of the Squared Residuals

Poll



If the point in the upper left corner of this scatterplot is removed from the data set, then what will happen to the slope of the regression (b) and to the correlation (r)?

- A) both will increase.
- B) both will decrease.
- C) b will increase, and r will decrease.
- D) b will decrease, and r will increase.
- E) both will remain the same.

The Effect of Outliers on the Line of Best Fit

- Outliers often have very large residual values
- The further a data point is from the rest of the data points (in the vertical direction), the larger its (squared) residual will be
- The mechanics that chooses the line of best fit tries to minimize the (squared) residual value(s) by attempting to get as close to the data point(s) as possible
- A data point with a large squared residual could adversely affect the slope of the regression line

Adding an Outlier to the Data

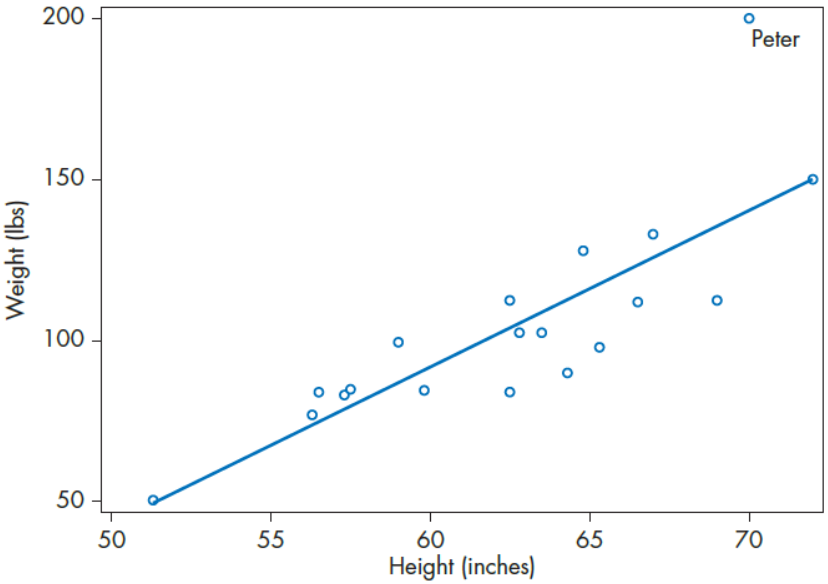


Figure 9.4: Regression Line for Height and Weight with Outlier Peter (70 inches, 200 lbs) Included

Predicted Weight = –200 + 4.9 x Height

The outlier has the effect of pulling the line up towards it making the line more steep with a slope of 4.9

Adding an Outlier to the Data

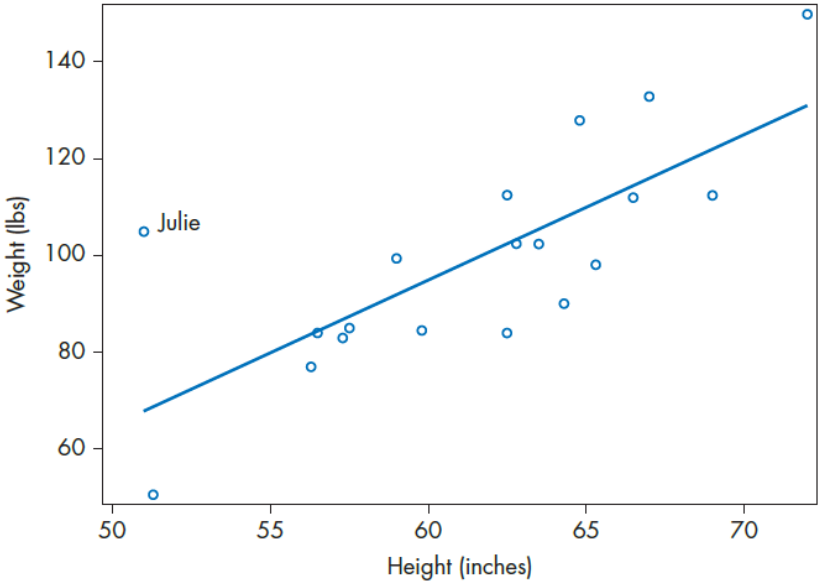


Figure 9.5: Regression Line for Height and Weight with Outlier Julie (52 inches, 105 lbs.) Included

Predicted Weight = –62.3 + 2.6 x Height

The outlier has the effect of pulling the line up towards it making the line less steep with a slope of 2.6

Adding an Outlier to the Data

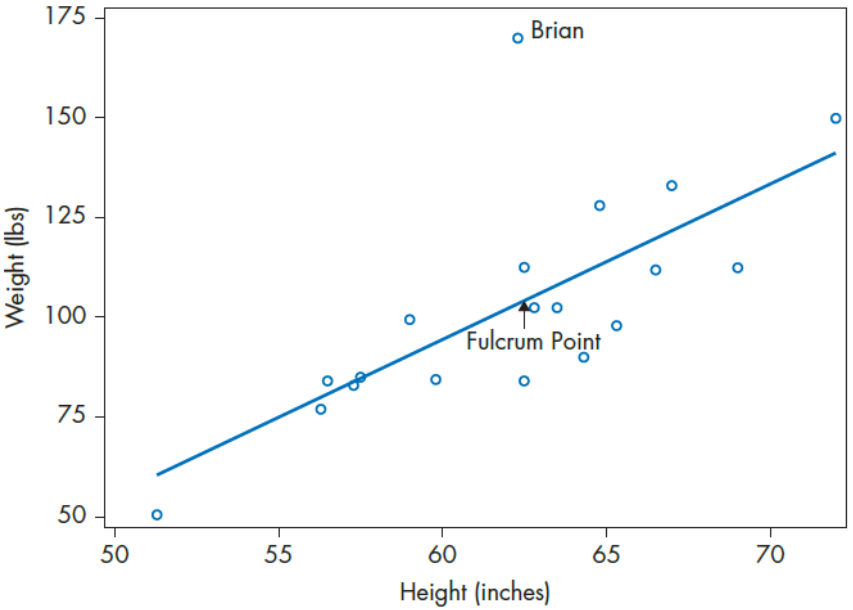


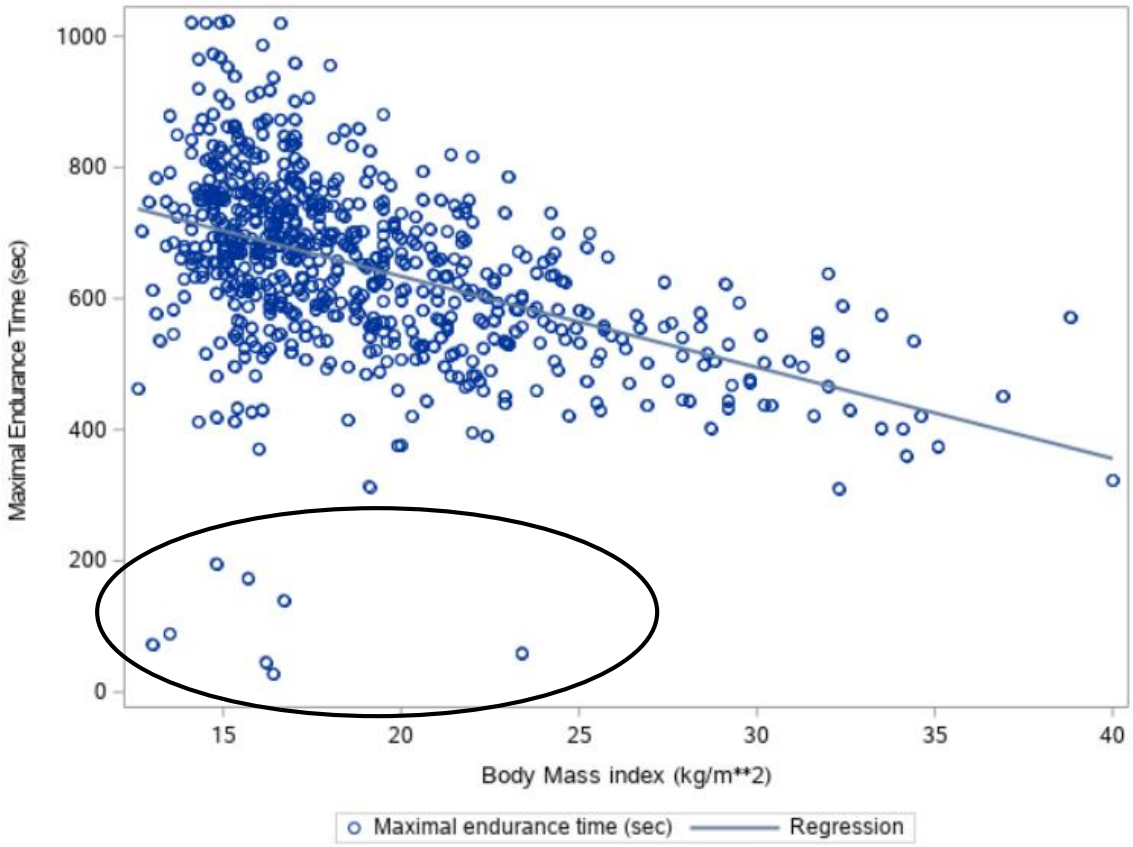
Figure 9.6: Regression Line for Height and Weight with Outlier Brian (62.3 inches, 170 lbs.) Included

Predicted Weight = –139 + 3.9 x Height

An outlier directly above the fulcrum point (mean height and mean weight) pulls the line towards it without change the slope of the line

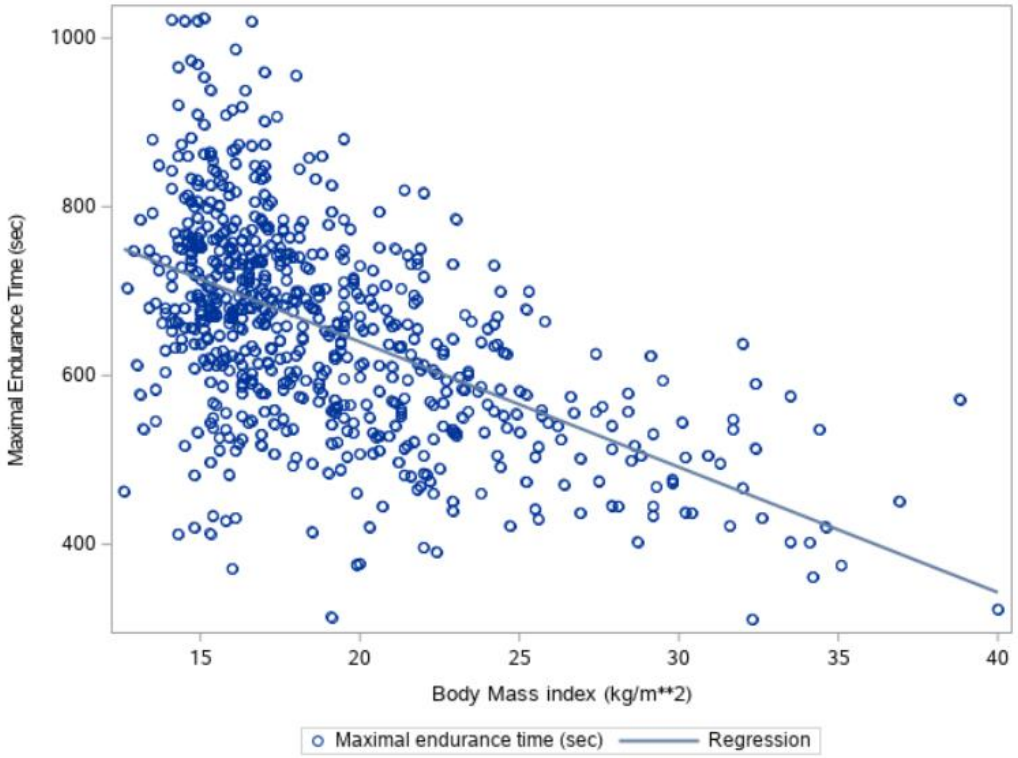
NHANES National Youth Fitness Survey

Predicted MET = 911.9 - 13.9 x BMI



Q. What do you think will happen to the slope of the line if we remove these points? Will the line get more steep or less steep?

NHANES National Youth Fitness Survey



Predicted MET = 936.96 – 14.85 x BMI

- Removing the outliers (children with METs less than 200 secs) increased the (negative) slope from -13.9 to -14.85)

Poll

If the point in the upper right corner of this scatterplot is removed from the data set, then what will happen to the slope of the regression (b) and to the correlation (r)?

- A) b will decrease, and r will increase
- B) b will remain the same, and r will increase.
- C) b will remain the same, and r will decrease.
- D) b will decrease, and r will remain the same.
- E) both will remain the same.

Residuals Diagnostics

- The linear model assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled.

Data* = *Model* + *Residual

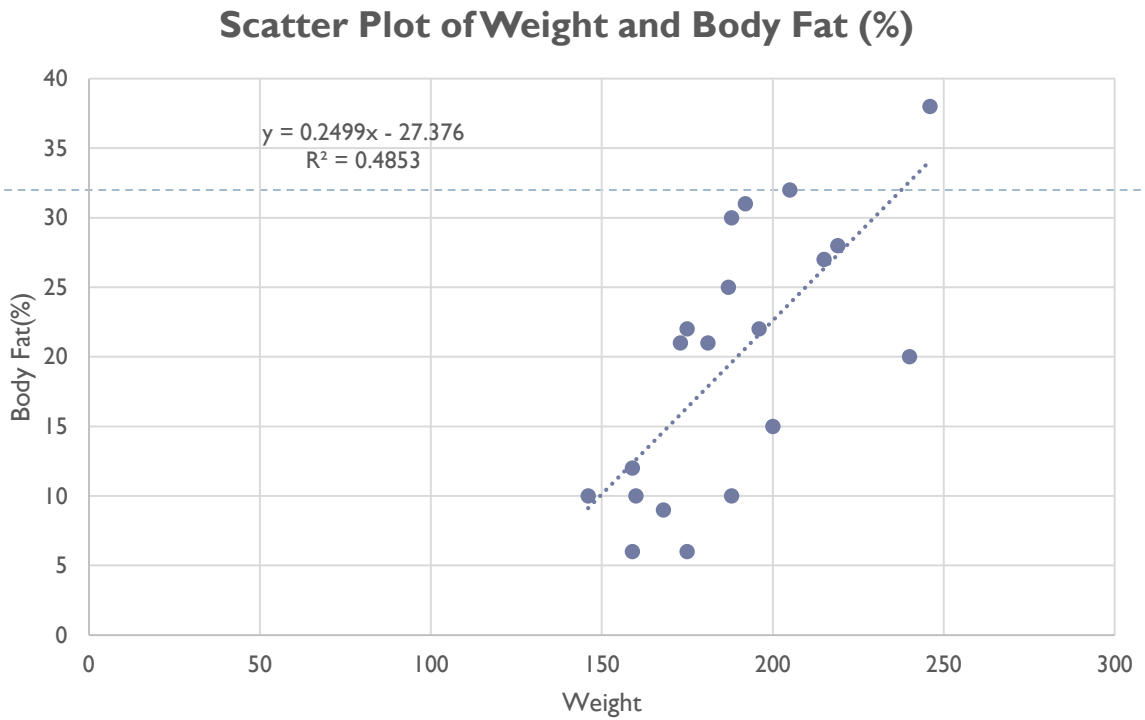
or (equivalently)

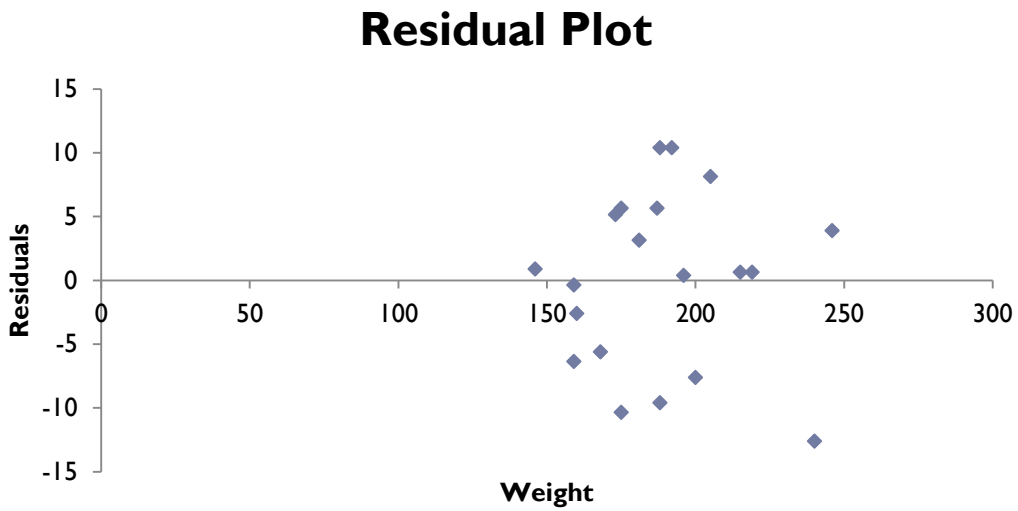
Residual* = *Data* – *Model

Or, in symbols, $e = y - \hat{y}$

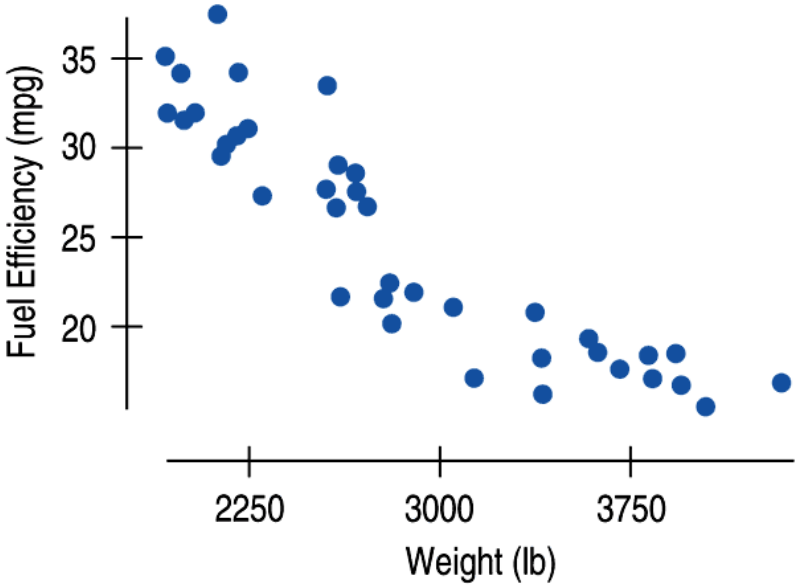
Residuals Diagnostics

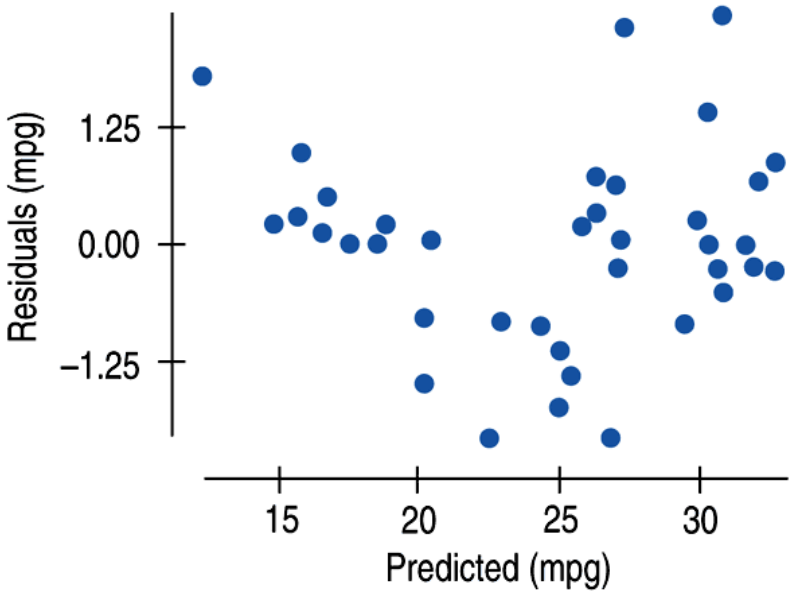
- Residuals help us to see whether the model makes sense.
- When a regression model is appropriate, nothing interesting should be left behind.
- After we fit a regression model, we usually plot the residuals in the hope of finding…nothing.


Scatter Plot of Weight and Body Fat (%)
y = 0.2499x - 27.376
R² = 0.4853
Body Fat(%)
Weight
0
5
10
15
20
25
30
35
40
0
50
100
150
200
250
300


Residual Plot
Residuals
Weight
15
10
5
0
-5
-10
-15
0
50
100
150
200
250
300

Fuel Efficiency Example





Standardized Residuals

- We can standardize each residual by subtracting the mean value (zero) and then dividing by the estimated standard deviation:

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

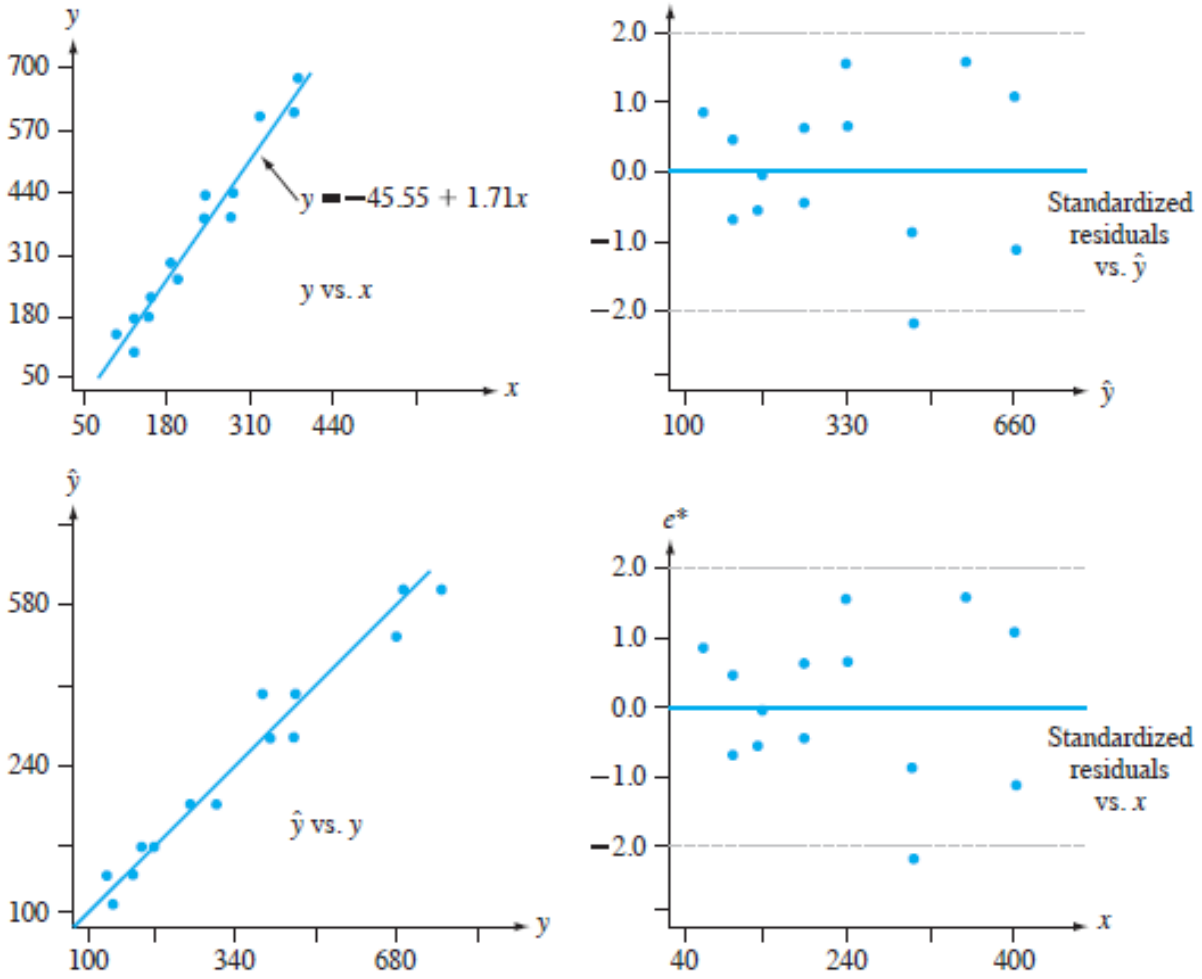
- We will denote e_i^* as the standardized residual

Diagnostic Plots

The basic plots that many statisticians recommend for an assessment of model validity and usefulness are the following:

1. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis
2. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis
3. $\hat{y}_i$ on the vertical axis versus y_i on the horizontal axis

Diagnostic Plots



Devore 7th Edition

Difficulties and Remedies

1. A nonlinear probabilistic relationship between x and y is appropriate.
2. The variance of ϵ (and of Y) is not a constant σ^2 but depends on x.
3. The selected model fits the data well except for a very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.