CHAPTER 12
REGRESSION DIAGNOSTICS: DETECTION OF MODEL VIOLATIONS

12-1 INTRODUCTION

12-2 THE STANDARD REGRESSION ASSUMPTION

13-3 INFLUENTIAL POINTS

11-4 MEASURES OF INFLUENCE

11-5 A STRATEGY FOR DEALING WITH PROBLEMATIC DATA POINTS

11-6 GRAPHS AFTER FITTING A MODEL

11-7 CHECKING STANDARD REGRESSION ASSUMPTION

INTRODUCTION

- This chapter presents methods for checking standard regression assumptions.
- We will rely mainly on graphical methods as opposed to applying rigid numerical rules to check for model violations.

THE STANDARD REGRESSION ASSUMPTIONS

- **1.** Assumptions about the form of the model (examining the **scatter plot**)

- **2.** Assumptions about the errors
  - The error ε_i, i = 1,2, ... , *n* has a **normal distribution**.
  - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have **mean zero**
  - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have the same (but unknown) variance σ^2 (as the **homogeneity** or the **homoscedasticity** assumption)
  - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, are independent of each other (their pairwise covariances are zero).

- **3. Assumptions about the observations**
  - All observations are equally **reliable** and have an approximately equal role in determining the regression results and in influencing conclusion

INFLUENTIAL POINTS

- **Overview:**
  - We learn about how data observations can potentially be influential in different ways.
  - If an observation has **a response value** that is very different from the predicted value based on a model, then that observation is called an outlier.
  - On the other hand, if an observation has a particularly unusual combination of **predictor values** (e.g., one predictor has a very different value for that observation compared with all the other data observations), then that observation is said to have high leverage.
  - Thus, there is a distinction between outliers and high leverage observations, and each can impact our regression analyses differently.
  - It is also possible for an observation to be both an outlier and have high leverage. Thus, it is important to know how to detect outliers and high leverage data points.
  - Once we've identified any outliers and/or high leverage data points, we then need to determine whether or not the points actually have an undue influence on our model.

INFLUENTIAL POINTS

- This section addresses all these issues using the following measures:
  - Leverages
  - Residuals
  - Studentized residuals (or internally studentized residuals)
  - (Unstandardized) deleted residuals (or PRESS prediction errors)
  - Studentized deleted residuals (or externally studentized residuals) [which Minitab calls deleted residuals]
  - Difference in fits (*DFFITS*)
  - Cook's distance measure

INFLUENTIAL POINTS

- Objective**:**
  - Upon completion of this section, you should be able to:
    - Understand the concept of an influential data point.
    - Know how to detect outlying *y* values by way of studentized residuals or studentized deleted residuals.
    - Understand leverage, and know how to detect outlying *x* values using leverages.
    - Know how to detect potentially influential data points by way of *DFFITS* and Cook's distance measure.

Distinction Between Outliers & High Leverage Observations

- In this section, we learn the distinction between outliers and high leverage observations. In short:
  - An outlier is a data point whose response y does not follow the general trend of the rest of the data.
  - A data point has high leverage if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values (e.g., with two predictors that are positively correlated, an unusual combination of predictor values might be a high value of one predictor paired with a low value of the other predictor).

Distinction Between Outliers & High Leverage Observations

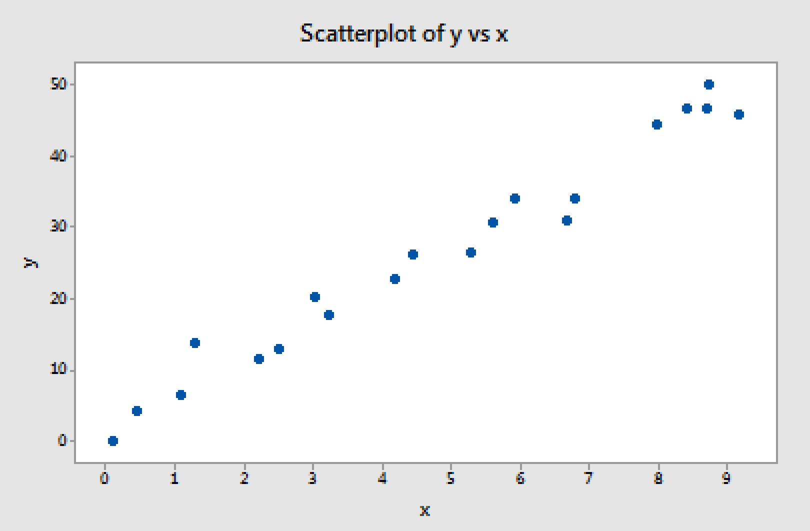
- Note that — for our purposes — we consider a data point to be an outlier *only if* it is extreme with respect to the other *y* values, not the *x* values.

- A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. Outliers and high leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential.

- One advantage of the case in which we have only one predictor is that we can look at simple scatter plots in order to identify any outliers and influential data points. Let's take a look at a few examples that should help to clarify the distinction between the two types of extreme values.

Example 1

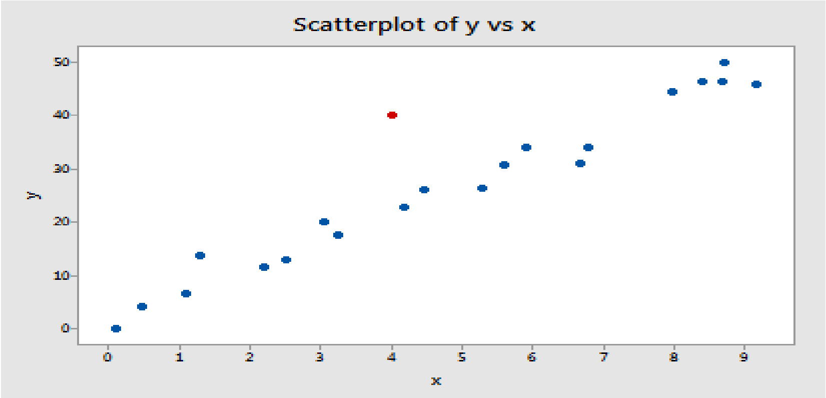
- Based on the definitions above, do you think the following **DataSet** contains any outliers? Or, any high leverage data points?



- All of the data points follow the general trend of the rest of the data, so there are no outliers (in the y direction). And, none of the data points are extreme with respect to x, so there are no high leverage points. Overall, none of the data points would appear to be influential with respect to the location of the best fitting line.

Example 2

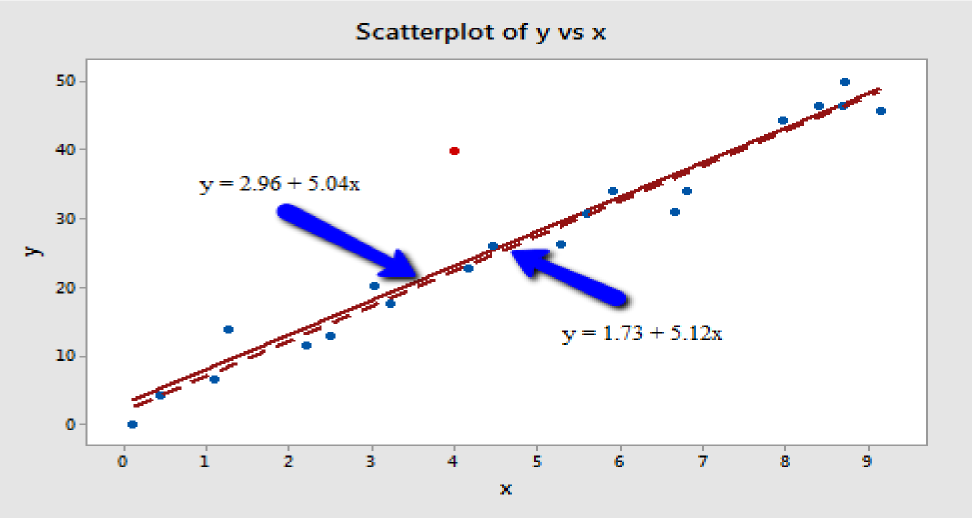
- Based on the definitions above, do you think the following **DataSet** contains any outliers? Or, any high leverage data points?



- Because the red data point does not follow the general trend of the rest of the data, it would be considered an outlier. However, this point does not have an extreme x value, so it does not have high leverage. Is the red data point influential? An easy way to determine if the data point is influential is to find the best fitting line twice — once with the red data point included and once with the red data point excluded.

Example 2

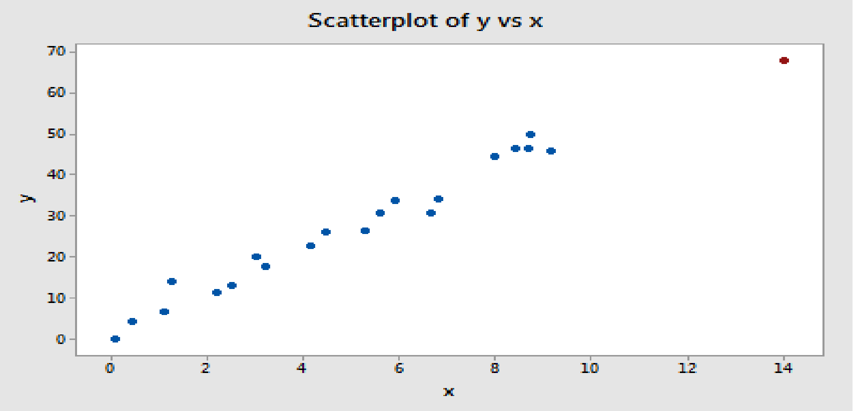
- The following plot illustrates the two best fitting lines:



- It's hard to even tell the two estimated regression equations apart! The solid line represents the estimated regression equation with the red data point included, while the dashed line represents the estimated regression equation with the red data point taken excluded. The slopes of the two lines are very similar — 5.04 and 5.12, respectively.

Example 3

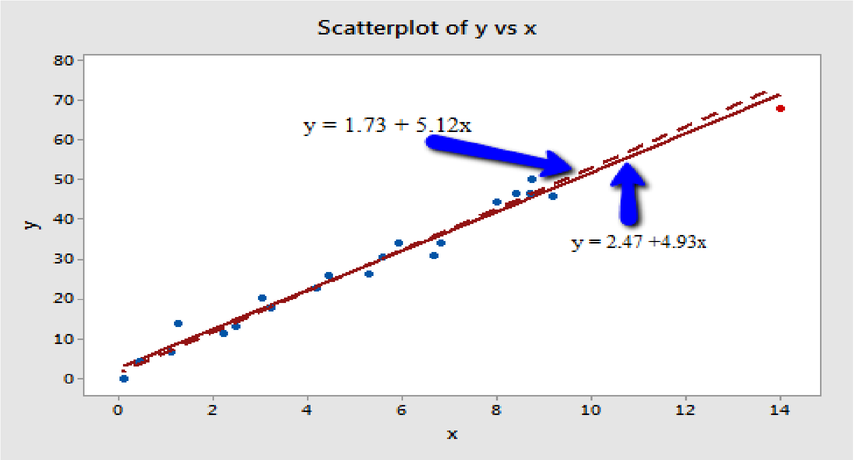
- Based on the definitions above, do you think the following **DataSet** contains any outliers? Or, any high leverage data points?



- In this case, the red data point does follow the general trend of the rest of the data. Therefore, it is *not* deemed an outlier here. However, this point does have an extreme x value, so it does have high leverage. Is the red data point influential? It certainly appears to be far removed from the rest of the data (in the x direction), but is that sufficient to make the data point influential in this case?

Example 3

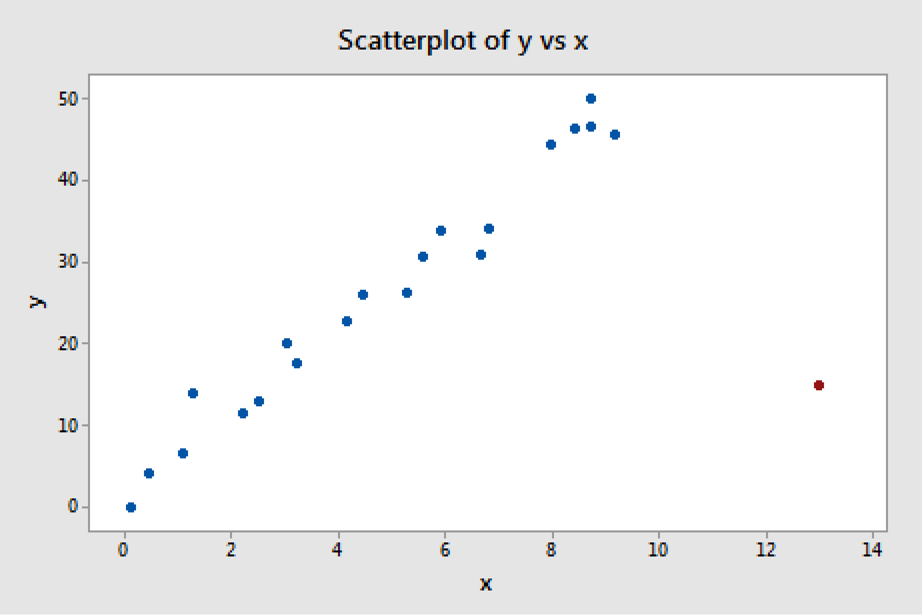
- The following plot illustrates two best fitting lines — one obtained when the red data point is included and one obtained when the red data point is excluded:



- It's hard to even tell the two estimated regression equations apart! The solid line represents the estimated regression equation with the red data point included, while the dashed line represents the estimated regression equation with the red data point taken excluded. The slopes of the two lines are very similar — 4.927 and 5.117, respectively.

Example 4

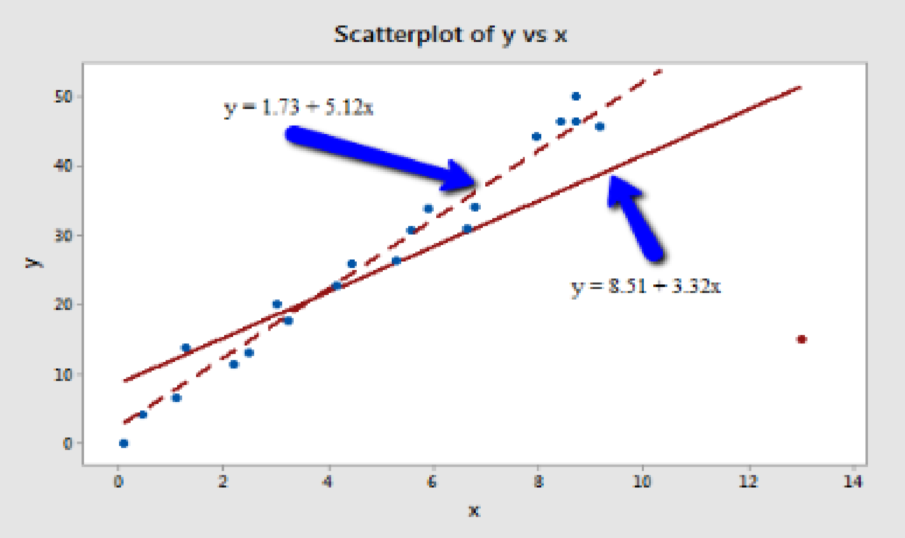
- Based on the definitions above, do you think the following **DataSet** contains any outliers? Or, any high leverage data points?



- In this case, the red data point is most certainly an outlier and has high leverage! The red data point does not follow the general trend of the rest of the data and it also has an extreme x value. And, in this case the red data point is influential.

Example 4

- The two best fitting lines — one obtained when the red data point is included and one obtained when the red data point is excluded:



- The solid line represents the estimated regression equation with the red data point included, while the dashed line represents the estimated regression equation with the red data point taken excluded. The existence of the red data point significantly reduces the slope of the regression line — dropping it from 5.117 to 3.320.

Summary

- The above examples — through the use of simple plots — have highlighted the distinction between outliers and high leverage data points. There were outliers in examples 2 and 4. There were high leverage data points in examples 3 and 4. However, only in example 4 did the data point that was both an outlier and a high leverage point turn out to be influential. That is, not every outlier or high leverage data point strongly influences the regression analysis. It is your job as a regression analyst to always determine if your regression analysis is unduly influenced by one or more data points.

- Of course, the easy situation occurs for simple linear regression, when we can rely on simple scatter plots to elucidate matters. Unfortunately, we don't have that luxury in the case of multiple linear regression. In that situation, we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both. Once we've identified such points we then need to see if the points are actually influential. We'll learn how to do all this in the next few sections!

Using Leverages to Help Identify Extreme x Values

- In this section, we learn about "**leverages**" and how they can help us identify extreme *x* values. We need to be able to identify extreme *x* values, because in certain situations they may highly influence the estimated regression function.

Using Leverages to Help Identify Extreme x Values

- **Definition and properties of leverages**:
  - You might recall from our brief study of the matrix formulation of regression that the regression model can be written succinctly as:

$$Y = X\beta + \epsilon$$

  - Therefore, the predicted responses can be represented in matrix notation as:

$$\hat{y} = Xb$$

  - And, if you recall that the estimated coefficients are represented in matrix notation as:

$$b = (X'X)^{-1}X'y$$

  - Then you can see that the predicted responses can be alternatively written as:

$$\hat{y} = X(X'X)^{-1}X'y$$

Using Leverages to Help Identify Extreme x Values

- **Definition and properties of leverages**:
  - That is, the predicted responses can be obtained by pre-multiplying the n × 1 column vector, y, containing the observed responses by the n × n matrix H:

$$H = X(X'X)^{-1}X'$$

  - That is:

$$\hat{y} = Hy$$

  - Do you see why statisticians call the n × n matrix **H "the hat matrix?"** That's right — because it's the matrix that puts the **hat "^"** on the observed response vector y to get the predicted response vector $\hat{y}$! And, why do we care about the hat matrix? Because it contains the "leverages" that help us identify extreme x values!
  - If we actually perform the matrix multiplication on the right side of this equation:

$$\hat{y} = Hy$$

Using Leverages to Help Identify Extreme x Values

- **Definition and properties of leverages**:
  - we can see that the predicted response for observation *i* can be written as a linear combination of the *n* observed responses y_1, y_2,…y_n:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \qquad \text{for } i = 1, \ldots, n$$

  - Where the weights h_{i1}, h_{i2},…h_{ii}…h_{in}: depend only on the predictor values.

$$\begin{aligned}
\hat{y}_1 &= h_{11}y_1 + h_{12}y_2 + \cdots + h_{1n}y_n \\
\hat{y}_2 &= h_{21}y_1 + h_{22}y_2 + \cdots + h_{2n}y_n \\
&\vdots \\
\hat{y}_n &= h_{n1}y_1 + h_{n2}y_2 + \cdots + h_{nn}y_n
\end{aligned}$$

  - Because the predicted response can be written as:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \qquad \text{for } i = 1, \ldots, n$$

Using Leverages to Help Identify Extreme x Values

- **Definition and properties of leverages**:
  - The **leverage**, h_{ii}, quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_{i.}$
  - That is, if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$.
  - On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$.
  - It's for this reason that the h_{ii} are called the "**leverages**."

Using Leverages to Help Identify Extreme x Values

- Here are some important properties of the leverages:
  - The leverage h_{ii} is a measure of the distance between the *x* value for the i^{th} data point and the mean of the *x* values for all *n* data points.
  - The leverage h_{ii} is a number between 0 and 1, inclusive.
  - The sum of the h_{ii} equals *p*, the number of parameters (regression coefficients including the intercept).

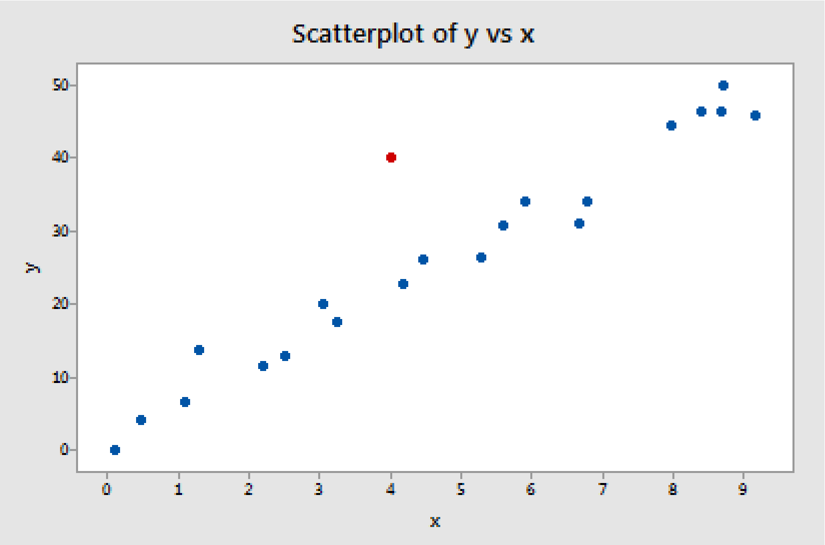
- The first bullet indicates that the leverage h_{ii} quantifies how far away the i^{th} x value is from the rest of the x values. If the i^{th} x value is far away, the leverage h_{ii} will be large; and otherwise not.

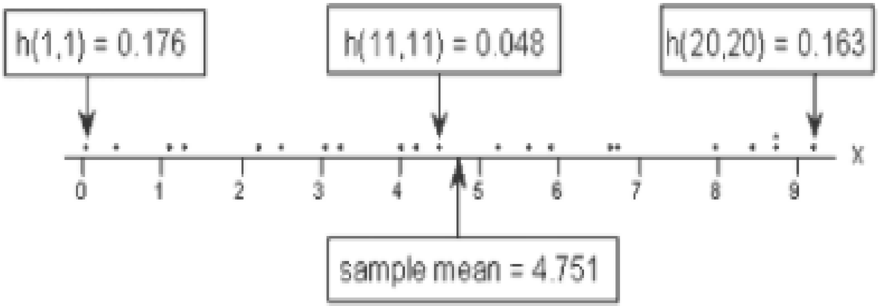
- Let's use the above properties — in particular, the first one — to investigate a few examples.

Example 2-Revisited

- Do any of the *x* values appear to be unusually far away from the bulk of the rest of the *x* values? Sure doesn't seem so, does it?

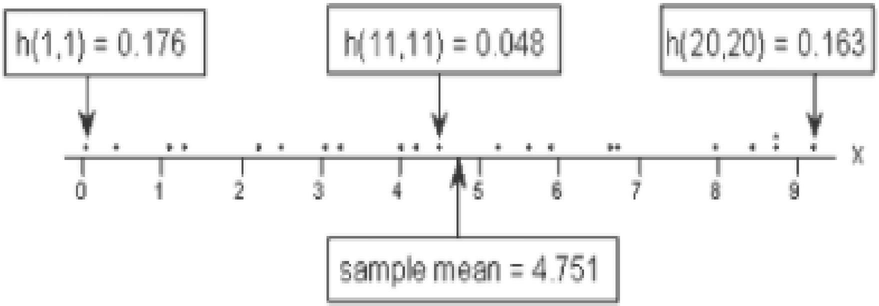
- Let's see if our intuition agrees with the leverages. Rather than looking at a scatter plot of the data, let's look at a scatter plot containing just the *x* values:

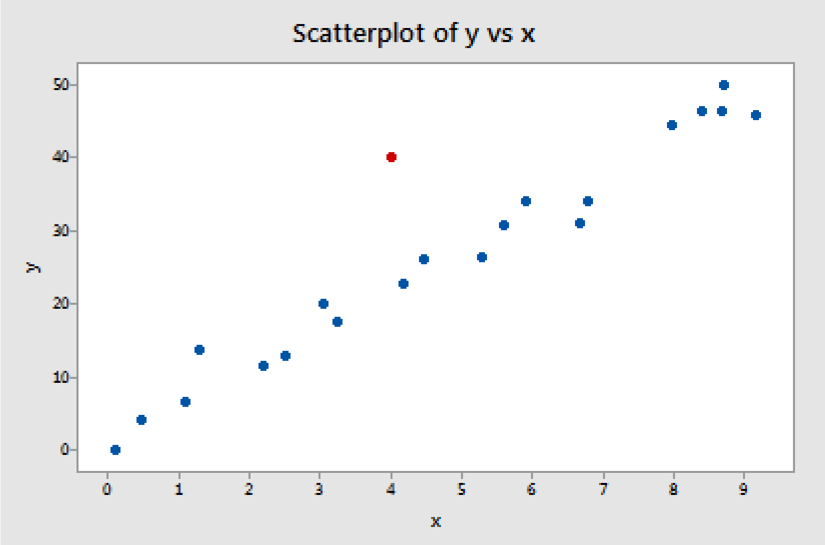


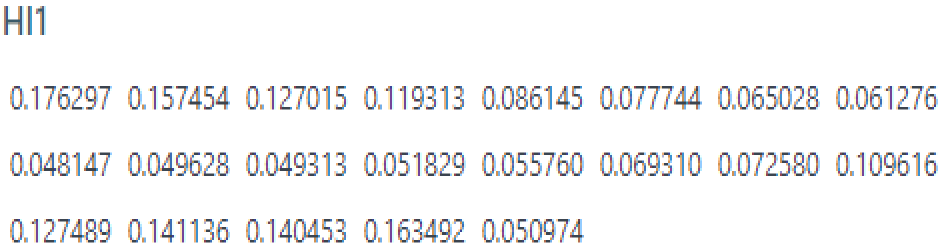


Example 2-Revisited

- Three of the data points — the smallest *x* value, an *x* value near the mean, and the largest *x* value — are labeled with their corresponding leverages. As you can see, the two *x* values furthest away from the mean have the largest leverages (0.176 and 0.163), while the *x* value closest to the mean has a smaller leverage (0.048). In fact, if we look at a sorted list of the leverages obtained in Minitab::





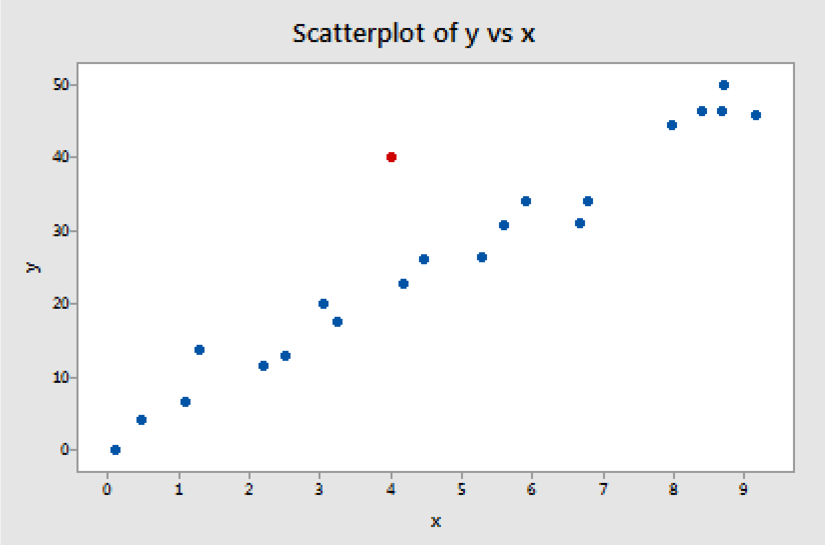


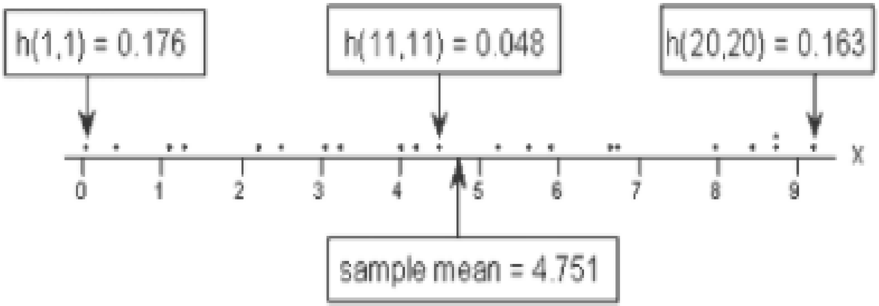
```
HI1

0.176297  0.157454  0.127015  0.119313  0.086145  0.077744  0.065028  0.061276

0.048147  0.049628  0.049313  0.051829  0.055760  0.069310  0.072580  0.109616

0.127489  0.141136  0.140453  0.163492  0.050974
```

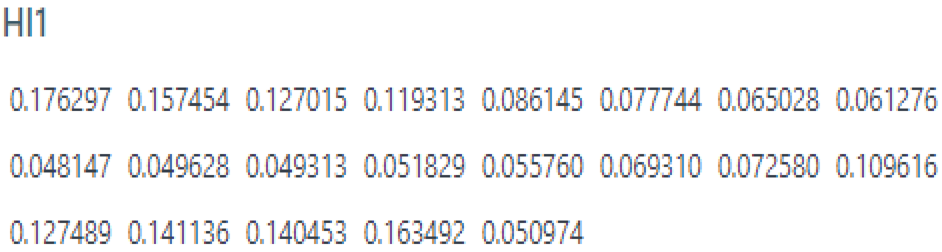
Example 2-Revisited

- We see that as we move from the small *x* values to the *x* values near the mean, the leverages decrease. And, as we move from the *x* values near the mean to the large *x* values the leverages increase again.

- You might also note that the sum of all 21 of the leverages add up to 2, the number of beta parameters in the simple linear regression model — as we would expect based on the third property mentioned above.







HI1

0.176297 0.157454 0.127015 0.119313 0.086145 0.077744 0.065028 0.061276

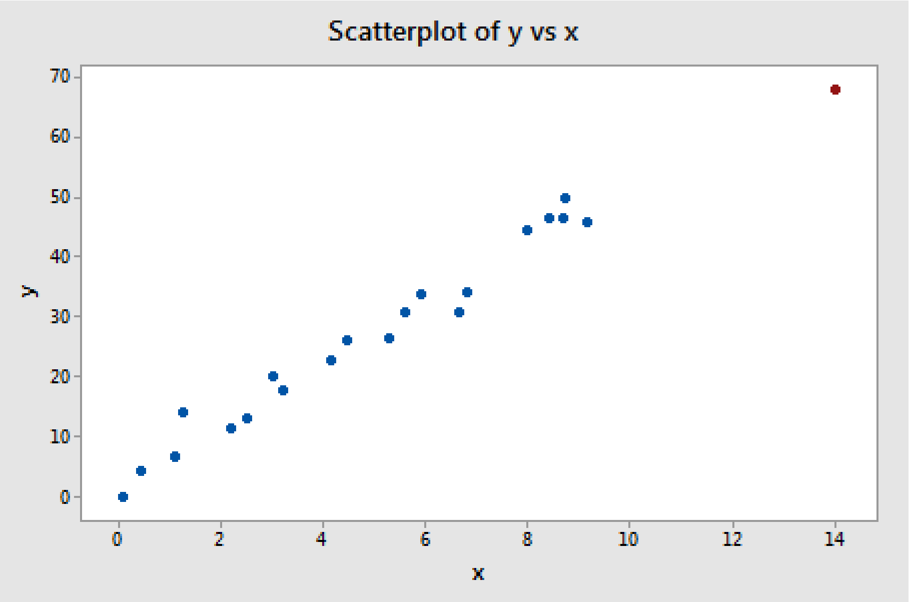
0.048147 0.049628 0.049313 0.051829 0.055760 0.069310 0.072580 0.109616

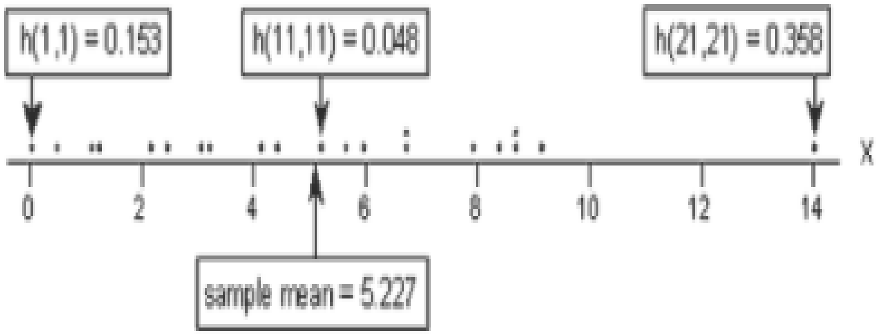
0.127489 0.141136 0.140453 0.163492 0.050974

Example 3-Revisited

- Do any of the *x* values appear to be unusually far away from the bulk of the rest of the *x* values? Hey, quit laughing! Sure enough, it seems as if the red data point should have a high leverage value. Let's see!?

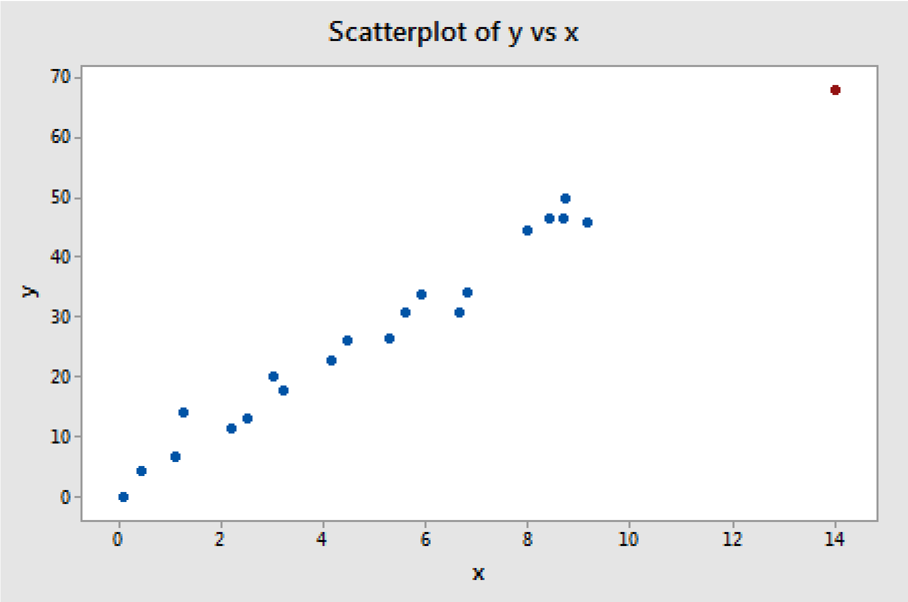
- A scatter plot containing just the *x* values:

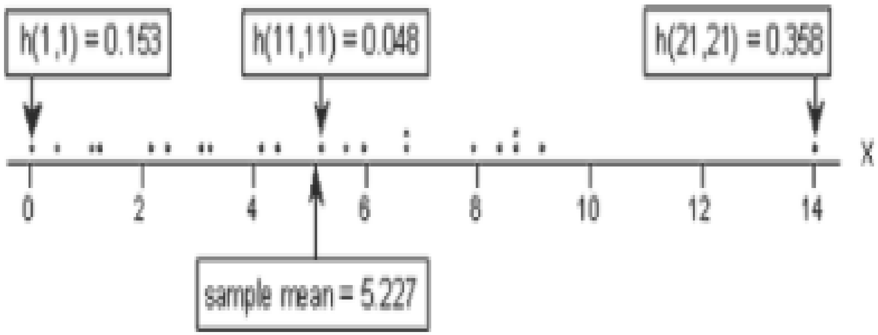


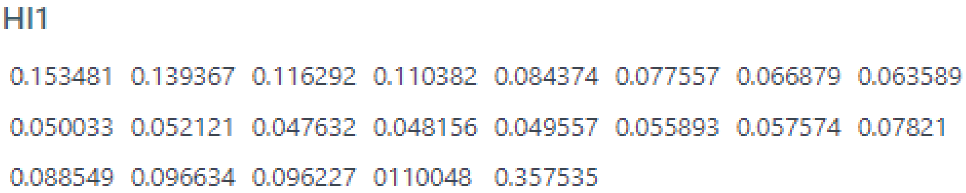


Example 3-Revisited

- Again, of the three labeled data points, the two *x* values furthest away from the mean have the largest leverages (0.153 and 0.358), while the *x* value closest to the mean has a smaller leverage (0.048). Looking at a sorted list of the leverages obtained in Minitab:





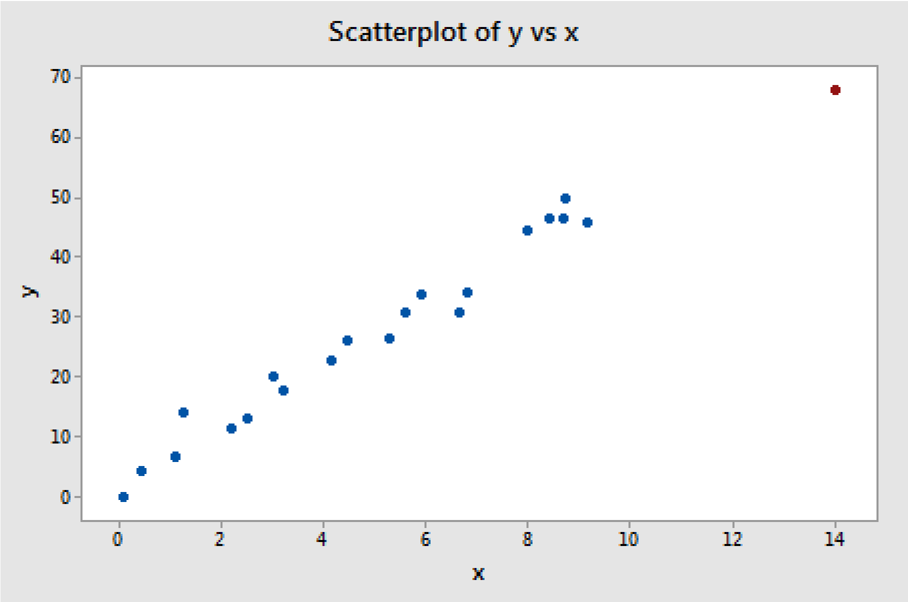


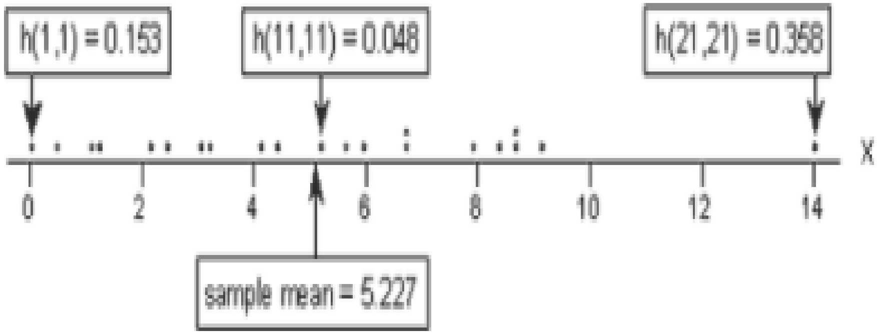
```
HI1

0.153481  0.139367  0.116292  0.110382  0.084374  0.077557  0.066879  0.063589

0.050033  0.052121  0.047632  0.048156  0.049557  0.055893  0.057574  0.07821

0.088549  0.096634  0.096227  0110048   0.357535
```

Example 3-Revisited

- We again see that as we move from the small *x* values to the *x* values near the mean, the leverages decrease. And, as we move from the *x* values near the mean to the large *x* values the leverages increase again. But, note that this time, the leverage of the *x* value that is far removed from the remaining *x* values (0.358) is much, much larger than all of the remaining leverages. This leverage thing seems to work!

- Don't forget to note again that the sum of all 21 of the leverages add up to 2, the number of beta parameters in the simple linear regression model. Again, we should expect this result based on the third property mentioned above.





```
HI1

0.153481  0.139367  0.116292  0.110382  0.084374  0.077557  0.066879  0.063589

0.050033  0.052121  0.047632  0.048156  0.049557  0.055893  0.057574  0.07821

0.088549  0.096634  0.096227  0110048  0.357535
```

Identifying data points whose x values are extreme

- The great thing about leverages is that they can help us identify *x* values that are extreme and therefore potentially influential on our regression analysis. How? Well, all we need to do is determine when a leverage value should be considered large. A common rule is to flag any observation whose leverage value, h_{ii}, is **more than 3 times larger** than the mean leverage value:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

- This is the rule that Minitab uses to determine when to flag an observation. That is, if:

$$h_{ii} > 3\left(\frac{p}{n}\right)$$

- then Minitab flags the observations as "Unusual X" (although it would perhaps be more helpful if Minitab reported "X denotes an observation whose X value gives it *potentially* large influence" or "X denotes an observation whose X value gives it large *leverage*").

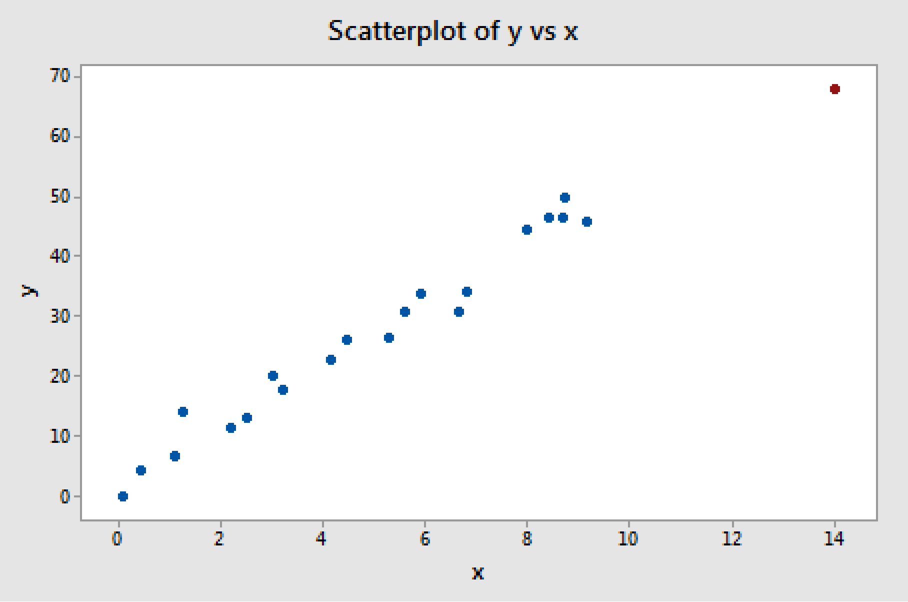
- As with many statistical "rules of thumb," not everyone agrees about this 3p/n cut-off and you may see 2p/n used as a cut-off instead. A refined rule of thumb that uses both cut-offs is to identify any observations with a leverage greater than 3p/n or, failing this, any observations with a leverage that is greater than 2p/n *and very isolated*.

Example 3-Revisited again

- Our intuition tells us that the red data point (*x* = 14, *y* = 68) is extreme with respect to the other *x* values. But, is the *x* value extreme enough to warrant flagging it? Let's see!:

- In this case, there are *n* = 21 data points and *p* = 2 parameters (the intercept β_0 and slope β_1). Therefore:

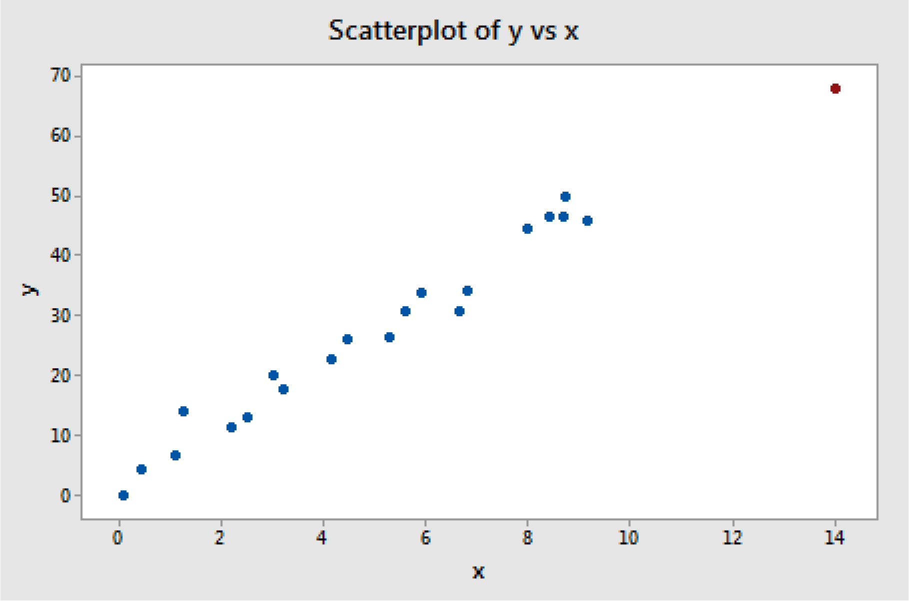
$$3\left(\frac{p}{n}\right) = 3\left(\frac{2}{21}\right) = 0.286$$

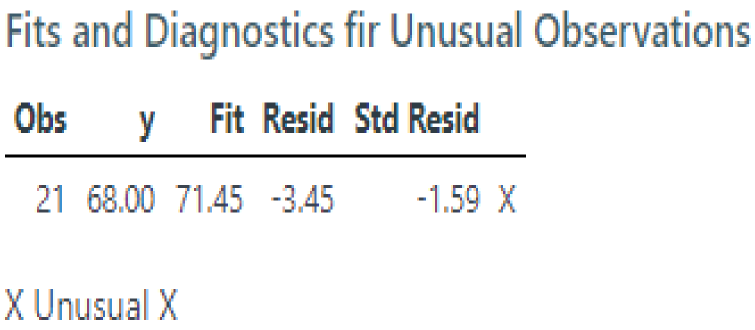


Example 3-Revisited again

- Now, the leverage of the data point — 0.358 (obtained in Minitab) — is greater than 0.286. Therefore, the data point should be flagged as having high leverage. And, that's exactly what Minitab does:

- **A word of caution!** Remember, a data point has large influence *only if* it affects the estimated regression function. As we know from our investigation of this data set in the previous section, the red data point does not affect the estimated regression function all that much. Leverages only take into account the extremeness of the x values, but a high leverage observation may or may not actually be influential.





Fits and Diagnostics fir Unusual Observations

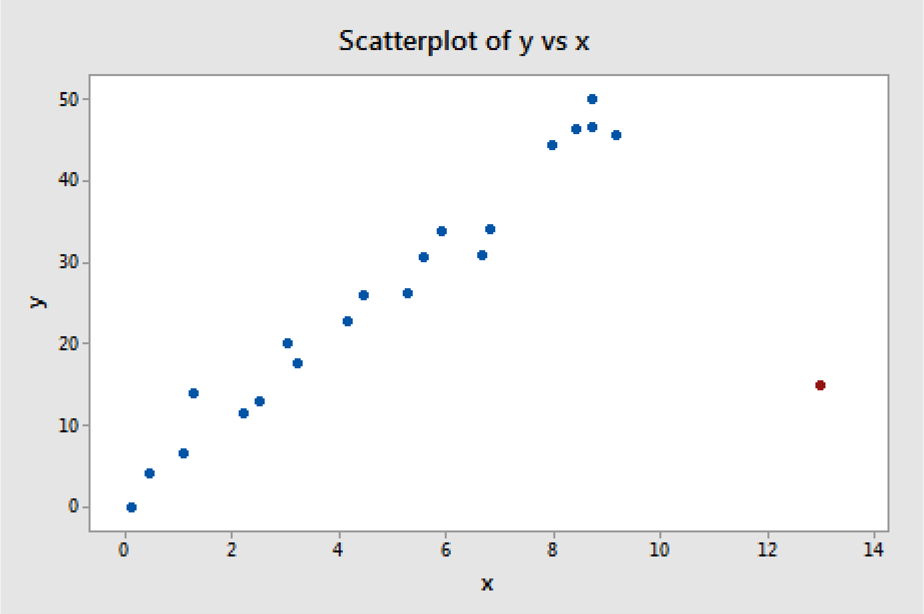
```
Obs      y    Fit  Resid  Std Resid
21   68.00  71.45  -3.45      -1.59  X
```

X Unusual X

Example 4-Revisited

- Our intuition tells us that the red data point (*x* = 13, *y* = 15) is extreme with respect to the other *x* values. Is the *x* value extreme enough to warrant flagging it?

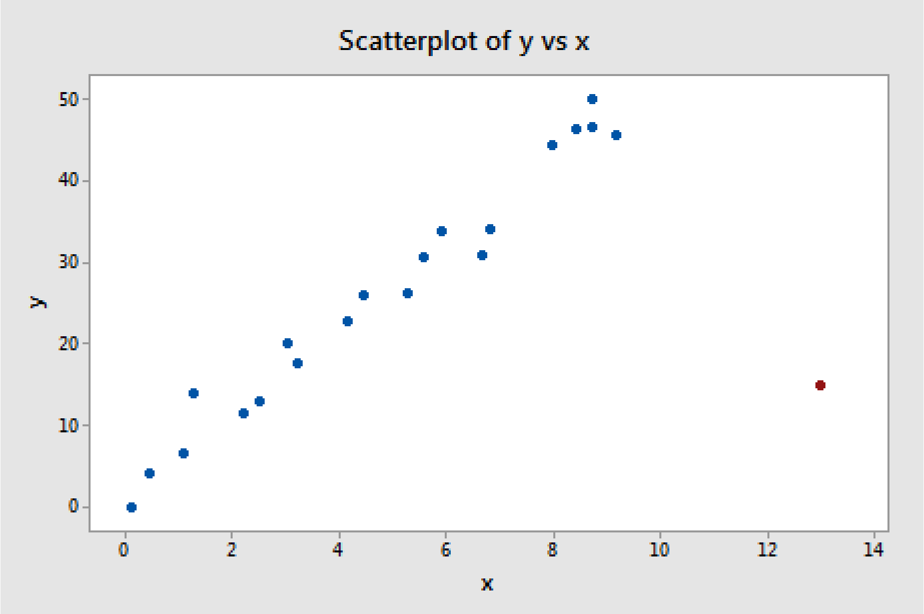
- Again, there are *n* = 21 data points and *p* = 2 parameters (the intercept β_0 and slope β_1). Therefore: $3\left(\frac{p}{n}\right) = 3\left(\frac{2}{21}\right) = 0.286$

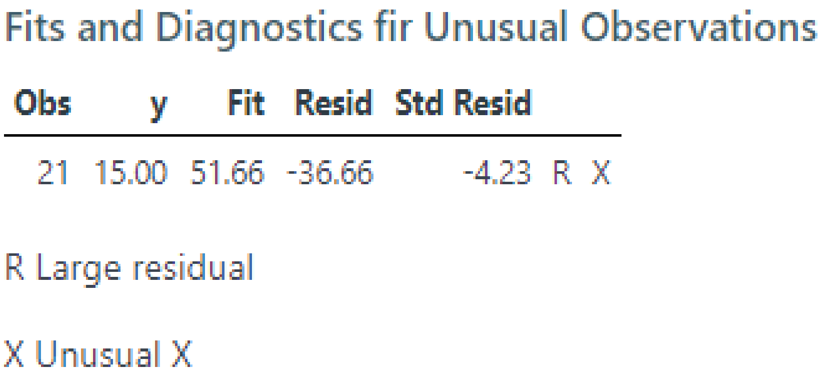


Example 4-Revisited

- Now, the leverage of the data point — 0.311 (obtained in Minitab) —is greater than 0.286. Therefore, the data point should be flagged as having high leverage, as it is:

- In this case, we know from our previous investigation that the red data point does indeed highly influence the estimated regression function. For reporting purposes, it would therefore be advisable to analyze the data twice — once with and once without the red data point — and to report the results of both analyses.





Fits and Diagnostics fir Unusual Observations

```
Obs      y     Fit   Resid  Std Resid
  21  15.00  51.66  -36.66      -4.23  R  X
```

R Large residual

X Unusual X

An important distinction

- There is such an important distinction between a data point that has high leverage and one that has high influence that it is worth saying it one more time:
  - The leverage merely quantifies the *potential* for a data point to exert strong influence on the regression analysis.
  - The leverage depends *only* on the predictor values.
  - Whether the data point is *influential* or not also depends on the observed value of the reponse

Identifying Outliers (Unusual y Values)

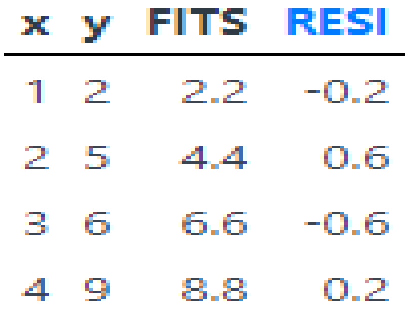
- Fitted values and *ordinary* residuals:
  - When fitting the linear model to a set of data by least squares, we obtain the fitted value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \cdots, n,$$

  - And the corresponding *ordinary* least squares residuals,

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n.$$

  - For example, consider the following very small (contrived) data set containing *n* = 4 data points (*x*, *y*).



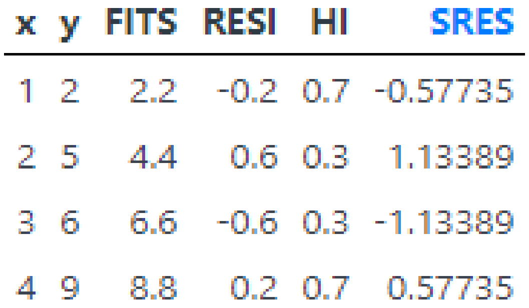
x	y	FITS	RESI
1	2	2.2	-0.2
2	5	4.4	0.6
3	6	6.6	-0.6
4	9	8.8	0.2

Identifying Outliers (Unusual y Values)

- **Studentized residuals (or internally studentized residuals)**:

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

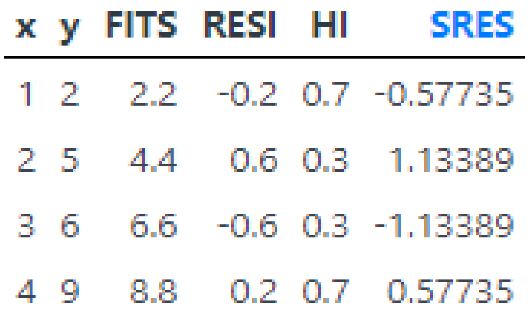
- Here, we see that the internally studentized residual for a given data point depends not only on the ordinary residual, but also the size of the mean square error (*MSE*) and the leverage h$_{ii}$:
- For example, consider again the (contrived) data set containing *n* = 4 data points (*x*, *y*):



x	y	FITS	RESI	HI	SRES
1	2	2.2	-0.2	0.7	-0.57735
2	5	4.4	0.6	0.3	1.13389
3	6	6.6	-0.6	0.3	-1.13389
4	9	8.8	0.2	0.7	0.57735

Identifying Outliers (Unusual y Values)

- For example, consider again the (contrived) data set containing *n* = 4 data points (*x*, *y*):



x	y	FITS	RESI	HI	SRES
1	2	2.2	-0.2	0.7	-0.57735
2	5	4.4	0.6	0.3	1.13389
3	6	6.6	-0.6	0.3	-1.13389
4	9	8.8	0.2	0.7	0.57735

- MSE=4.0

$$r_1 = \frac{-0.2}{\sqrt{0.4(1-0.7)}} = -0.57735$$

$$r_2 = \frac{0.6}{\sqrt{0.4(1-0.3)}} = 1.13389$$

Identifying Outliers (Unusual y Values)

- The good thing about internally studentized residuals is that they quantify how large the residuals are in standard deviation units, and therefore can be easily used to identify outliers:
  - An observation with an internally studentized residual that is larger than 3 (in absolute value) is generally deemed an outlier. (Sometimes, the term "outlier" is reserved for an observation with an externally studentized residual that is larger than 3 in absolute value—we consider externally studentized residuals in the next section.)
  - Recall that Minitab flags any observation with an internally studentized residual that is larger than 2 (in absolute value).

Identifying Outliers (Unusual y Values)

➢ Deleted Residuals:

- So far, we have learned various measures for identifying extreme x values (high leverage observations) and unusual y values (outliers).
- When trying to identify outliers, one problem that can arise is when there is a potential outlier that influences the regression model to such an extent that the estimated regression function is "pulled" towards the potential outlier, so that it isn't flagged as an outlier using the standardized residual criterion.
- To address this issue, **deleted residuals** offer an alternative criterion for identifying outliers. The basic idea is to delete the observations one at a time, each time refitting the regression model on the remaining n–1 observations.
- Then, we compare the observed response values to their fitted values based on the models with the i^{th} observation deleted. This produces **(unstandardized) deleted residuals**.
- Standardizing the deleted residuals produces **studentized deleted residuals**, also known as **externally studentized residuals**.

Identifying Outliers (Unusual y Values)

- **Deleted Residuals:**
  - If we let:
    - y_i denote the observed response for the ith observation, and
    - $\hat{y}_{(i)}$ denote the predicted response for the ith observation based on the estimated model with the ith observation deleted
  - Then the ith (unstandardized) deleted residual is defined as:

$$d_i = y_i - \hat{y}_{(i)}$$

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

Identifying Outliers (Unusual y Values)

- **Deleted Residuals:**
  - Why this measure? Well, data point i being influential implies that the data point "pulls" the estimated regression line towards itself. In that case, the observed response would be close to the predicted response. But, if you removed the influential data point from the data set, then the estimated regression line would "bounce back" away from the observed response, thereby resulting in a large deleted residual. That is, a data point having a large deleted residual suggests that the data point is influential.

Identifying Influential Data Points

- In this section, we learn the following two measures for identifying influential data points**:**
  - **Difference in fits (DFFITS)**
  - **Cook's distance**

- The basic idea behind each of these measures is the same, namely to delete the observations one at a time, each time refitting the regression model on the remaining n–1 observations. Then, we compare the results using all n observations to the results with the i^{th} observation deleted to see how much influence the observation has on the analysis. Analyzed as such, we are able to assess the potential impact each data point has on the regression analysis.

Difference in fits (DFFITS)

- The difference in fits for observation *i*, denoted DFFITS$_i$, is defined as**:**

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

- An observatio ntial if the absolute value of its *DFFITS* value is greater than:

$$2\sqrt{\frac{p+1}{n-p-1}}$$

- Where as always *n* = the numbe ations and *p* = the number of parameters including the intercept. It is important to keep in mind that this is not a hard-and-fast rule, but rather a guideline only! It is not hard to find different authors using a slightly different guideline. Therefore, I often prefer a much more subjective guideline, such as a data point is deemed influential if the absolute value of its *DFFITS* value sticks out like a sore thumb from the other *DFFITS* values. Of course, this is a qualitative judgment, perhaps as it should be, since outliers by their very nature are subjective quantities.

Using Cook's distance measures

- The beauty of the above examples is the ability to see what is going on with simple plots. Unfortunately, we can't rely on simple plots in the case of multiple regression. Instead, we must rely on guidelines for deciding when a Cook's distance measure is large enough to warrant treating a data point as influential.
- Here are the guidelines commonly used:
  - If D_i is greater than 0.5, then the i^{th} data point is worthy of further investigation as it **may be influential**.
  - If D_i is greater than 1, then the i^{th} data point is **quite likely to be influential**.
  - If D_i sticks out like a sore thumb from the other D_i values, it is **almost certainly influential**.

Cook's distance

- Just jumping right in here, Cook's distance measure, denoted D_i, is defined as:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE}\left(\frac{h_{ii}}{(1 - h_{ii})^2}\right)$$

- It looks a little messy, but the main thing to recognize is that Cook's D_i depends on both the residual, e_i , and the leverage, h_{ii}. That is, both the x value and the y value of the data point play a role in the calculation of Cook's distance.

- In short:
  - D_i directly summarizes how much *all* of the fitted values change when the i^{th} observation is deleted.
  - A data point having a large D_i indicates that the data point strongly influences the fitted values.

WHAT TO DO WITH THE OUTLIER

- Identifying and handling outliers and influential data points is a "wishy-washy" business
- Use the measures described herein only as a way of screening their data set for *potentially* influential data points.
- Here are the recommended strategies for dealing with problematic data points:
  - First, check for obvious data errors:
    - If the error is just a data entry or data collection error, correct it.
    - If the data point is not representative of the intended study population, delete it.
    - If the data point is a procedural error and invalidates the measurement, delete it.

WHAT TO DO WITH THE OUTLIER

- Here are the recommended strategies for dealing with problematic data points:
  - Consider the possibility that you might have just misformulated your regression model::
    - Did you leave out any important predictors?
    - Should you consider adding some interaction terms?
    - Is there any nonlinearity that needs to be modeled?

A Strategy for Dealing with Problematic Data Points

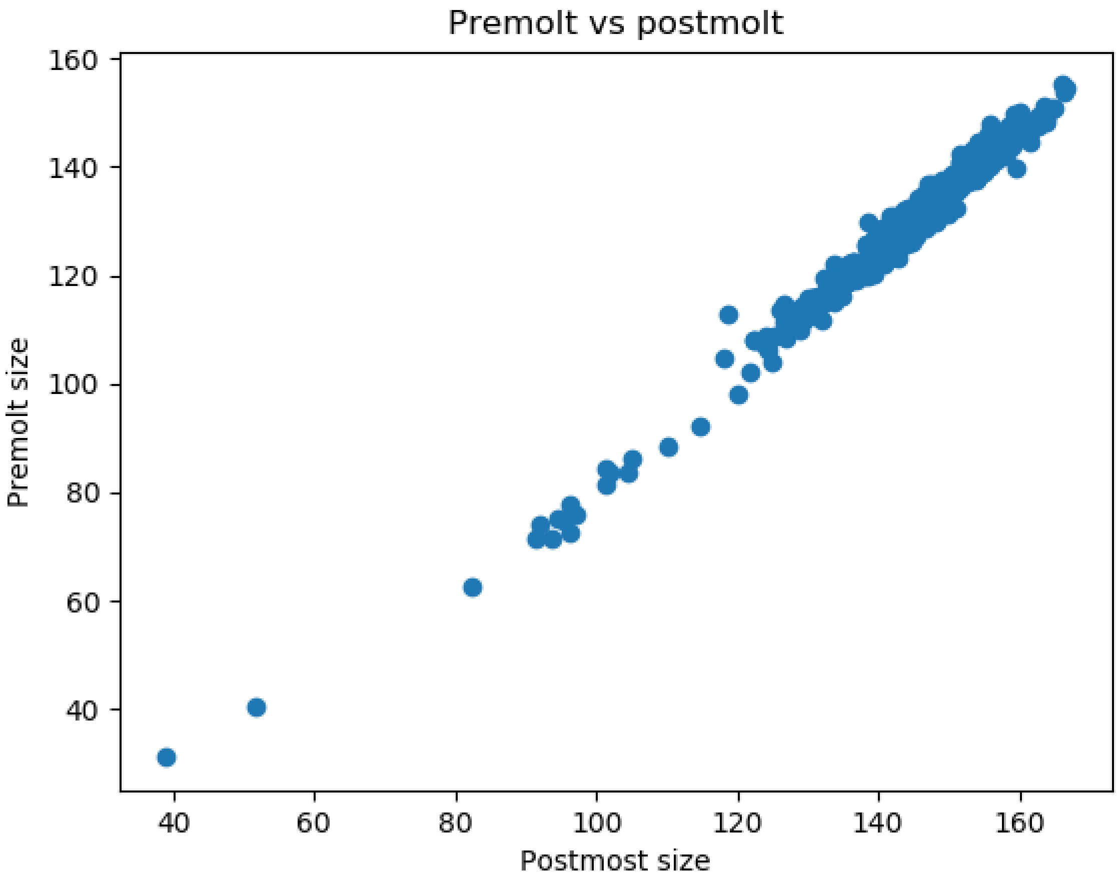
- Here are the recommended strategies for dealing with problematic data points:
  - If nonlinearity is an issue, one possibility is to just reduce the scope of your model. If you do reduce the scope of your model, you should be sure to report it, so that readers do not misuse your model.
  - Decide whether or not deleting data points is warranted: Did you leave out any important predictors?
    - Do not delete data points just because they do not fit your preconceived regression model.
    - You must have a good, objective reason for deleting data points.
    - If you delete any data after you've collected it, justify and describe it in your reports.
    - If you are not sure what to do about a data point, analyze the data twice — once with and once without the data point — and report the results of both analyses.

GRAPHS AFTER FITTING A MODEL

- The graphs after fitting a model to the data help in checking the assumptions and in assessing the adequacy of the fit of a given model. These graphs can be grouped into the following classes:
  - Graphs for checking the linearity and normality assumptions
  - Graphs for the detection of outliers and influential observations
  - Diagnostic plots for the effect of variables

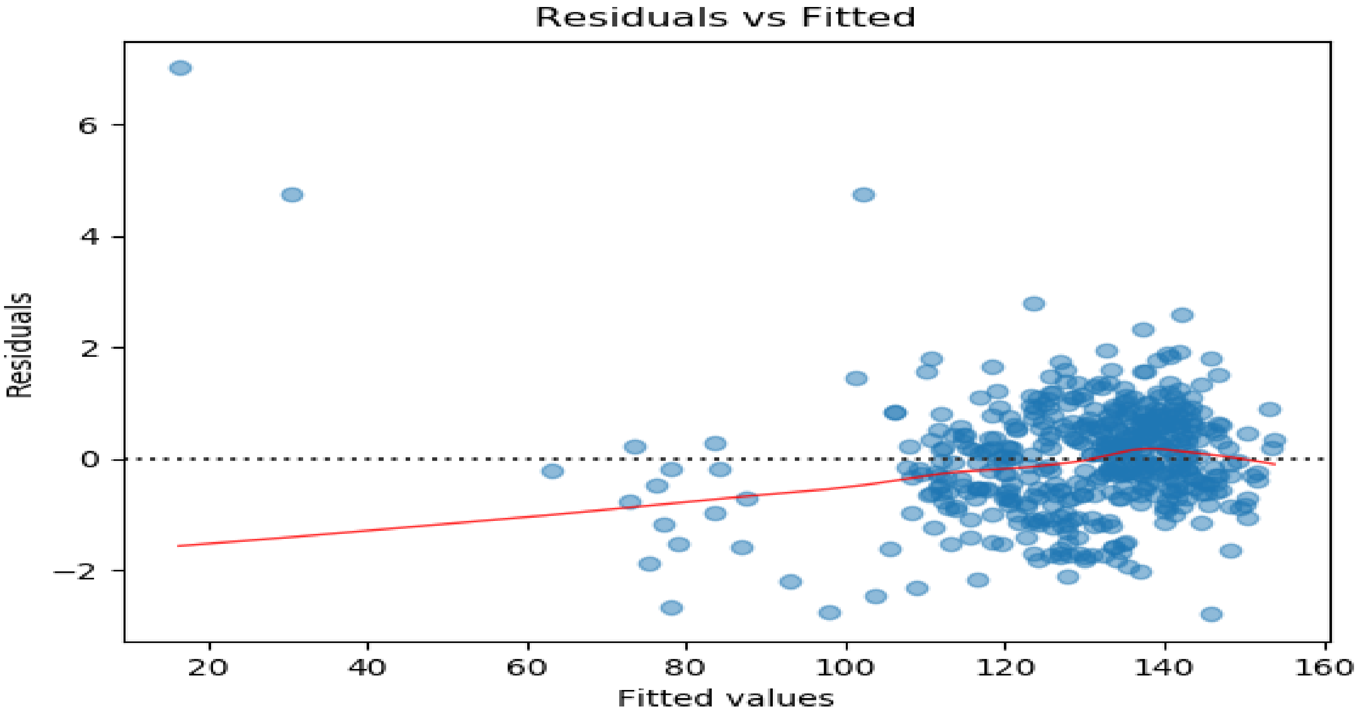
Checking assumptions about the form of the model

- We visualize the relationship between our variables to get an intuitive grasp of the data.



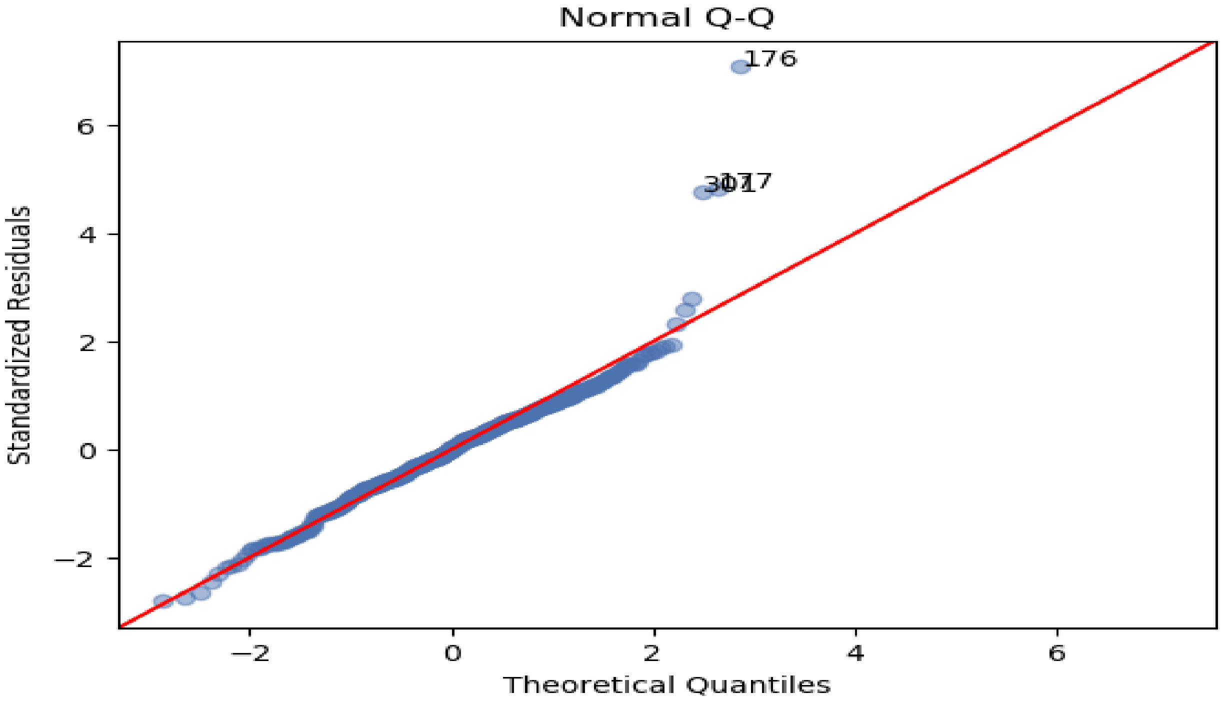
Checking assumptions about the errors

- The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have **mean zero**.
  - We use residual value vs fitted value plot



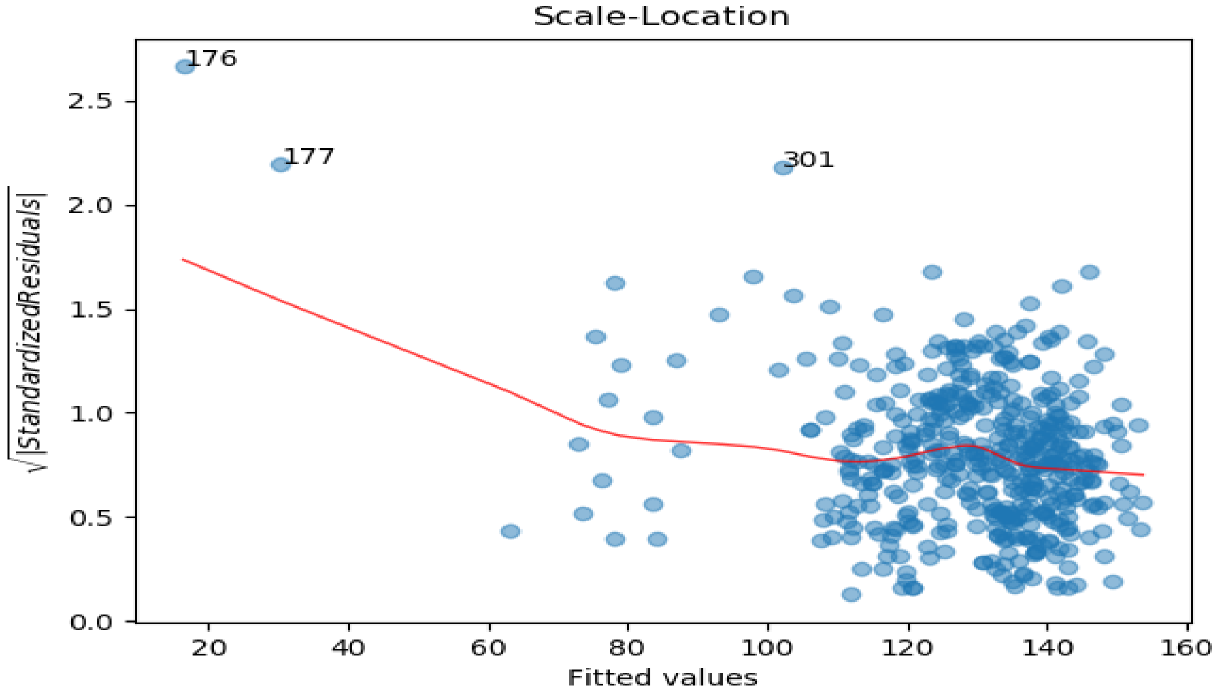
Checking assumptions about the errors

- The error ε_i, i = 1,2, ... , *n* has a **normal distribution**..
  - We use normal Q-Q plot



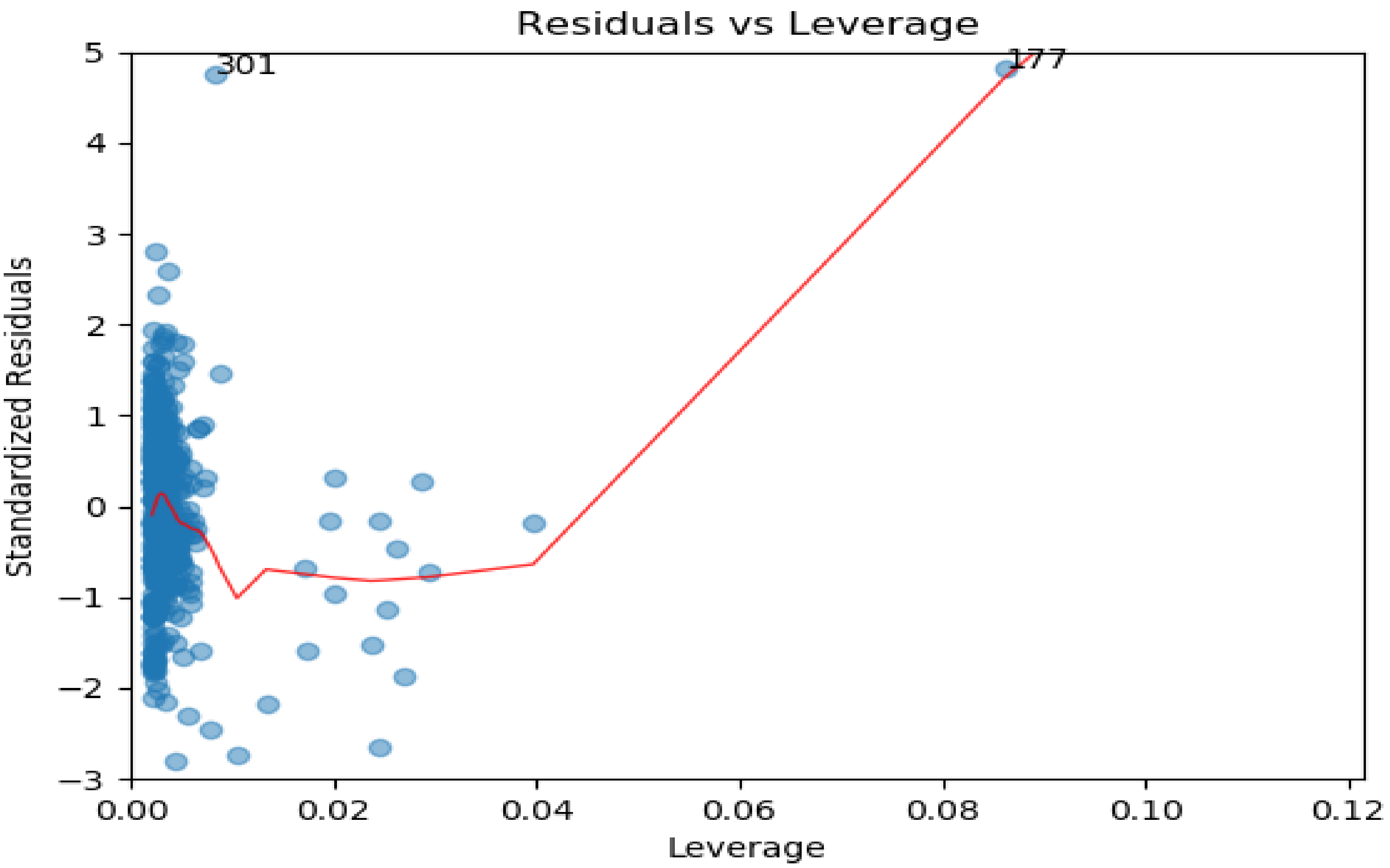
Checking assumptions about the errors

- The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have the same (but unknown) variance σ^2 (as the **homogeneity** or the **homoscedasticity** assumption)
  - – We use scale-location plot



Graphs for the detection of outliers and influential observations

➢


Residuals vs Leverage
Standardized Residuals
Leverage