Multiple Linear Regression & Diagnostics

Lab 3

Review

Multiple Linear Regression Assumptions

In multiple linear regression, we assume that some linear combination of covariates $x_{i1}, \ldots, x_{ip}$ correctly characterizes the true mean of each Y_i, like this

$$E[Y_i | x_{i1}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

We also assume that the natural variations ("errors") of Y_i's around that underlying mean line are independently distributed as $\epsilon_i \sim \text{Normal}(0, \sigma^2)$ for some $\sigma^2 > 0$.

L inearity of the mean: The mean of Y_i is an unknown but linear function of the covariates.

I ndependent: Data from each observation are independent of each other.

N ormality: The distribution of the Y_i's around their mean is Normal.

E quality of variance: The variability of the Y_i's around their mean is always the same regardless of the values of any covariates x_i.

Confounding and Effect Modification

A variable is a **confounding variable** if it satisfies two conditions:

1. it is a risk factor for the outcome.
2. it is associated with the exposure, but not a consequence of the exposure.

If an analysis adjusting for a confounding variable gives an appreciably different result than an unadjusted analysis, we informally say the added variable is **confounding** the exposure-outcome association and prefer the adjusted analysis. Failure to control for confounding can lead to *bias*.

Effect modification is a change in the exposure-outcome association across levels of an **effect modifier**. It is a *property* of the true exposure-outcome association, and exists (or not) independent of your study design or analysis.

Indicator Variables and ANOVA

An **indicator** or **dummy** variable takes values 0 or 1. We can represent indicator variables as **indicator functions**. For example, a linear regression model of age categories $([0, 25), [25, 50), [50, \infty))$ to predict total cholestrol (`tc`) can be written as:

$$E[\text{tc}_i | \text{age}_i] = \beta_0 + \beta_1 I(25 \leq \text{age}_i < 50) + \beta_2 I(\text{age}_i \geq 50),$$

where

$$I(25 \leq \text{age}_i < 50) = \begin{cases} 1 & \text{if } 25 \leq \text{age}_i < 50 \\ 0 & \text{if age}_i < 25 \text{ or age}_i \geq 50 \end{cases}$$

Analysis of Variance (ANOVA) explains the variance of outcome as a function of one or more categorical factors. ANOVA computes the variances explained by these factors, and tests for effects and/or

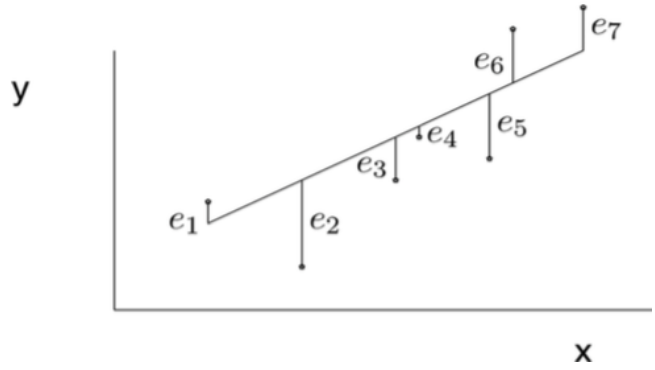
interactions. ANOVA is basically equivalent to linear regression using indicator variables. For example, if we run an ANOVA model for total cholesterol against age categories (defined as above), we will obtain the total variance explained by all age categories, which is equivalent to the sum of variances explained by each category reported in the linear regression model.

Diagnostics

After we fit a model to our data, we must check to see that the **LINE** assumptions are reasonable assumptions for our data. Two evaluations that we use are *residual evaluation* and *outlier detection.*

Residuals

Recall from lecture that a **residual** e_i is the difference between the observed value and the predicted mean: $e_i = Y_i - \widehat{Y}_i$.[1] In the case of simple linear regression, we're able to visualize residuals:



Residual plots can help us to

- detect outliers (= observations with very large residuals; see next section)
- look for nonlinear trends (the **L** in **LINE**)
- assess Normality (the **N** in **LINE**)
- assess homoscedasticity (the equal variance assumption, the **E** in **LINE**)

Types of Residuals[2]

Residuals: $e_i = Y_i - \widehat{Y}_i$

Also known as raw residuals. These estimate the unobserved error in our regression model.

Standardized residuals: $z_i = \dfrac{e_i}{s_e} = \dfrac{e_i}{\sqrt{SSE/(N-p-1)}}$

Raw residual divided by the estimated standard deviation. More useful because the range of values similar to $N(0,1)$.

(Externally) Studentized residuals: $r_{(i)} = \dfrac{e_i}{s_{e(i)}\sqrt{(1-h_i)}} = \dfrac{e_i}{\widehat{\sigma_{(i)}}\sqrt{1-h_i}}$

Also known as jackknife residuals or leave-one-out residuals. Here h_i is

[1] Sometimes the residual is denoted as $\widehat{\varepsilon}_i$ because ε_i is our theoretical deviation or 'error,' $Y_i - E[Y_i]$, and $e_i = \widehat{\varepsilon}_i$ is the deviation we observe based on $\hat{Y}_i$, our model's estimate of $E[Y_i]$.

[2] There are different definitions of standardized and studentized residuals, here we use the definition consistent with the lecture notes.

the *leverage*, discussed below, and $\widehat{\sigma_{(i)}}$ is the estimated standarded deviation with i^{th} observation removed. Preferred by some statisticians because they are more robust to outliers.

(Internally) Studentized residuals: $r_i = \dfrac{e_i}{s_e\sqrt{(1-h_i)}} = \dfrac{e_i}{\widehat{\sigma}\sqrt{1-h_i}}$
This follows t_{N-p-1}, and is also close to $N(0,1)$ when the sample size is relatively large.

Notice that when the sample size is relatively large, standardized and studentized residuals will be very similar.

Residual Plots

There are several diagnostic residual plots we can look at to evaluate our **LINE** assumptions.

- ⋆ *Histogram of standardized residuals*
  Should appear Normally distributed.
- ⋆ *Histogram of studentized or jackknife residuals*
  Will roughly approximate a t distribution (or Normal when N is large).
- ⋆ *Boxplot*
  Residuals should be symmetrical without too many residuals beyond the whiskers.
- ⋆ *Q-Q plot*
  Residuals should follow a straight line.
- ⋆ *Scatter plot against either $\widehat{Y}_i$ or x_{ij}*
  Should be a fluffy, patternless cloud (i.e., no cone shape, no linear trend, no curvature).

Autocorrelation in Residuals

Autocorrelation refers to correlation between residuals across different cases. This may happen if the cases are time dependent (e.g. time series data). One way to assess the presence of autocorrelation is the Durbin-Watson statistic:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2}$$

We can test for autocorrelation under the following hypothesis:

$$\begin{cases} H_0 : d = 2 & \text{null hypothesis of no autocorrelation} \\ H_1 : d \neq 2 & \text{alternative hypothesis of either positive or negative autocorrelation} \end{cases}$$

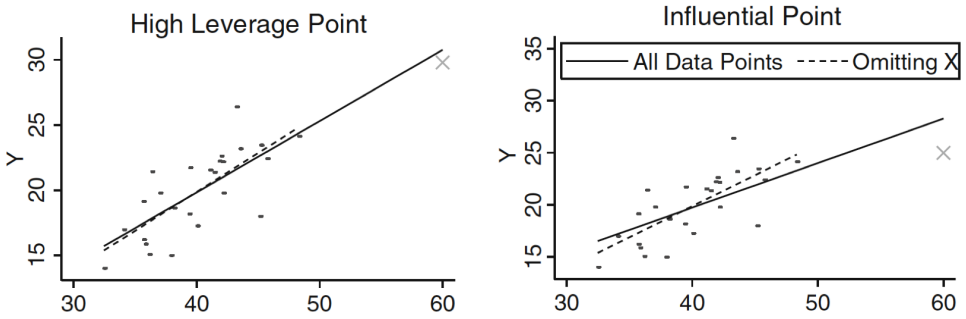
.

Outlier Detection

Outlying data points have the potential to greatly affect the slope of our fitted regression line. We can calculate several quantities for each data point to assess which ones either potentially or actually exert strong influence on our fitted model.

Leverage

The **leverage** of the ith observation is denoted with h_i, and it measures how far observation i's covariates are from the overall covariate average. In the case of simple linear regression, we're able to visualize leverage:



Thus we see that high leverage does not always mean that the outlier is influential, or 'bad'. It just means that the point has *potential* to exert strong influence on our model fit. The "rule of thumb" is that if $h_i > \frac{2(p+1)}{n}$ then i is a high leverage point, where p is the number of covariates (excluding the intercept) and n is the number of observations.

Cook's Distance

The **Cook's distance** of the ith observation is denoted by d_i, and it measures how much the regression parameters would change if observation i is deleted. Unlike leverage, Cook's distance reflects the actual amount of influence that a data point has on the model fit. A Cook's distance value is considered large if $d_i > \frac{4}{n-2}$.

DFFITS

Similar to Cook's distance, **DFFITS** is a measure of how influential a point is to the fit of a line. It represents the difference between $\widehat{y_i}$ and $\widehat{y_{(-i)i}}$ where $\widehat{y_i}$ is the estimated outcome of data point i using a linear model estimated with the full data set and $\widehat{y_{(-i)i}}$ represents the estimated outcome of data point k using a linear model estimated with a data set missing data point i. Points with $|\text{DFFITS}_i| > 2\sqrt{\frac{p+1}{n}}$ should be investigated. DFFITS values should be close or identical to Cook's distance since they are conceptually similar.

DFBETA

Whereas the previous quantities assessed each point's impact on overall model fit, the $\textbf{DFBETA}_{ik}$ statistic quantifies how much a particular coefficient β_k changes if we delete a particular observation i. The number is presented in standard error units of how far the "new" β_k falls from the "old" β_k, so the rule of thumb is that $|\text{DFBETA}_{ik}| > \frac{2}{\sqrt{n}}$ will detect about the top 5% of the influential data points.

Dealing with Outliers

If you observe outliers, first go back and check the original data to be sure there were no errors in data entry. You may want to fit the model with and without one or a small number of outliers, to see how much the model changes due to these outliers. However, be very careful of reporting results with outliers deleted. This could make your model overfitted to the original data. You might instead report your inferences both with and without the outliers. If things don't change substantially, report results using all of the data.

Ultimately, removing outliers should be the last resort in a data analysis. The existence of many outliers is often an indication that either linear model assumptions are violated, there is a problem with data

collection, or the data itself requires a different or more complex model. However, remember that because many of these diagnostics rely on quantiles, there may be random chance that a certain percentage of data points "lie outside" acceptable standard deviations. If circumstances do require a removal of outliers, a thorough explanation of reasoning must also be included in the analysis.

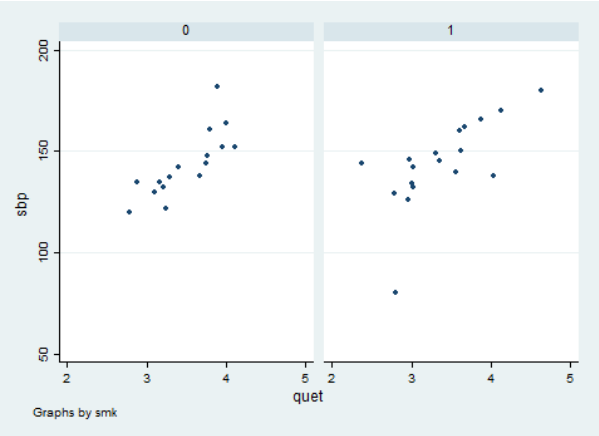
Data analysis in Stata

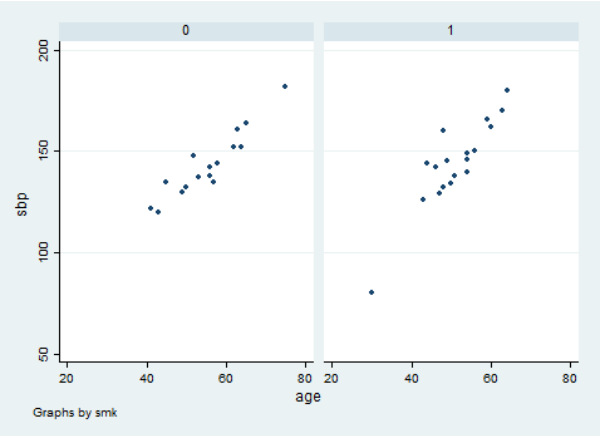
Our dataset lab03_outliers.dta contains the systolic blood pressure (`sbp`), body size (quetelet index, `quet`), age and smoking history (`smk=0` if nonsmoker and `smk=1` if current or previous smoker) for a sample of 34 people. Our response variable is `sbp`.

Initial analysis

Let's first plot the outcome `sbp` by each covariate (`quet` and `age`) stratified by smoking history.

```
. scatter sbp quet, by(smk)
. scatter sbp age, by(smk)
```





There appears to be an outlier in the `smk=1` group in the scatter plot of `sbp` versus `quet` and the scatter plot of `sbp` versus `age`. We'll keep an eye on this observation (which is person #34).

Let's use Stata to regress `sbp` on `smk`, `quet`, and `age`.

```
 regress sbp smk quet age
```

```
      Source |       SS           df       MS      Number of obs   =        34
-------------+----------------------------------   F(3, 30)        =     41.20
       Model |  9634.06305         3  3211.35435   Prob > F        =    0.0000
    Residual |  2338.55459        30  77.9518198   R-squared       =    0.8047
-------------+----------------------------------   Adj R-squared   =    0.7851
       Total |  11972.6176        33  362.806595   Root MSE        =     8.829

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         smk |   9.491079   3.158924     3.00   0.005     3.039696    15.94246
        quet |   2.769163   4.827272     0.57   0.570    -7.089441    12.62777
         age |   1.898326   .2843106     6.68   0.000     1.317686    2.478965
       _cons |    28.1944   11.42189     2.47   0.019     4.867779    51.52102
------------------------------------------------------------------------------
```

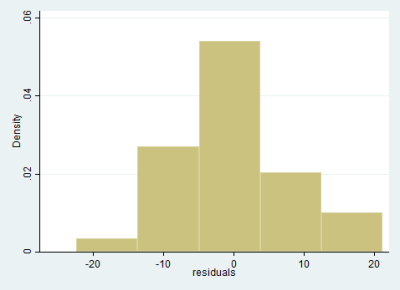
Residuals

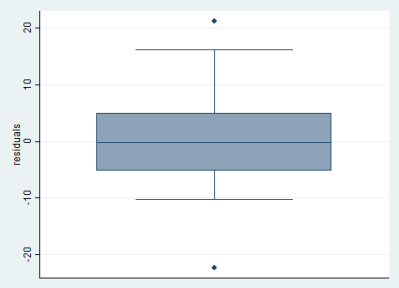
From our previous model fit we can calculate the model's residuals.

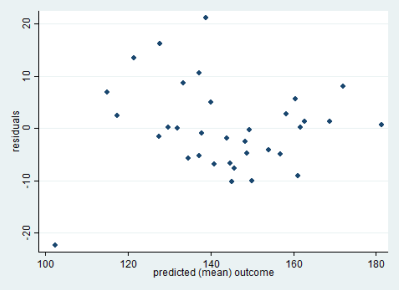
```
* generate a new variable named 'res', that contains the residuals
predict res, residuals
label variable res "Residuals"
```

Now that we've calculated the residuals, let's look at a histogram, a boxplot, and a scatter plot of residuals versus fitted values.

```
* generate a new variabled named 'sbp_hat', that contains the predicted means
predict sbp_hat,
label variable sbp_hat "predicted (mean) outcome"
hist res
graph box res
scatter res sbp_hat
```







Question Evaluate these plots. What are we looking for in each plot? Do you see any indication that our model assumptions are not being met? Do the raw residuals have mean zero? Are you surprised?

The histogram appears to be Normally distributed; the boxplot appears to be symmetric; but the scatter plot appears to have a parabolic shape, and there is a far outlier. Therefore, we may wonder if the linearity assumption holds.

We may want to add non-linear terms. But here let's look at the outliers first.

We can run a one-sample t-test for: $H_0 : \mathbb{E}[residuals] = 0$ vs. $H_1 : \mathbb{E}[residuals] \neq 0$.

We can also verify analytically that the residuals should have mean zero, according to theories of linear regression.

Here is some math (you don't have to know this, just in case you are interested):

For simple linear equation:

$$\mathbb{E}[y_i|x_i] = \beta_0 + \beta_1 x_i$$

We find the least squares estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ by:

$$\begin{aligned}
C(\beta_0, \beta_1) &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) \stackrel{set}{=} 0 \\
\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \stackrel{set}{=} 0
\end{aligned}$$

Solve the equations to find $\hat{\beta}_0$ and $\hat{\beta}_1$. Since they are solutions to these equations, we can plug them into the first equation and get:

$$\begin{aligned}
-2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n}\hat{y}_i = 0 \\
\Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{y}_i = 0
\end{aligned}$$

Therefore, the mean of the residuals is given as

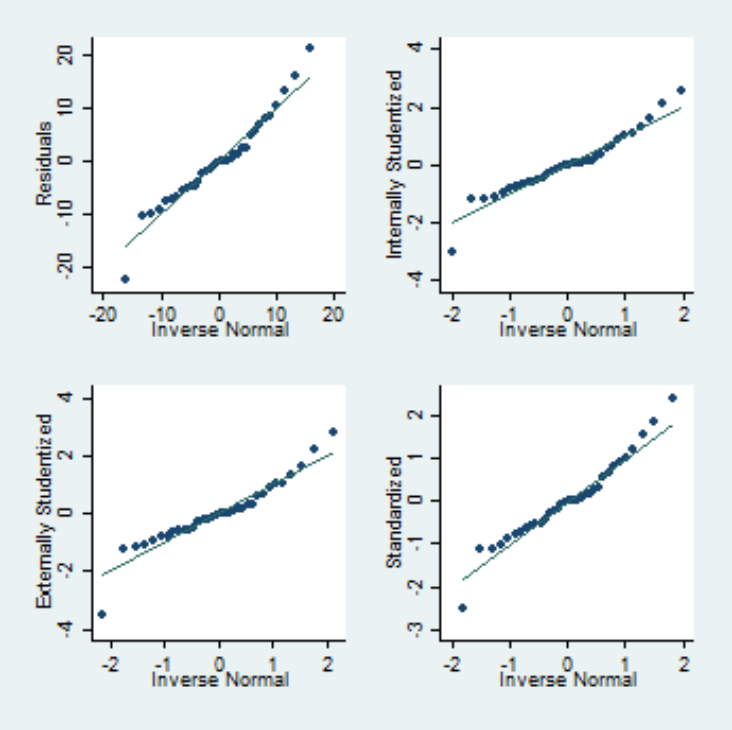
$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} e_i &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{y}_i\right) \\
&= 0
\end{aligned}$$

In order to formally assess whether this dataset contains any outliers, let's compute and visualize the unstandardized, standardized, internally studentized, and externally studentized residuals using Normal Q-Q plots. [3]

```
* generate a new variable named 'inStu', that contains internal studentized residuals
predict inStu, rstandard
label variable inStu "Internally Studentized"
* generate a new variable named 'exStu', that contains externally studentized residuals
predict exStu, rstudent
label variable exStu "Externally Studentized"
* generate a new variable named 'stdRes' that standardizes the residuals
gen stdRes = res / e(rmse)
label variable stdRes "Standardized"

* create Normal Q-Q plots of the residuals
qnorm res, saving(qqres, replace)
qnorm inStu, saving (qqinStu, replace)
qnorm exStu, saving(qqexStu, replace)
qnorm stdRes, saving (qqstd, replace)

*combine into one plot
gr combine qqres.gph qqinStu.gph qqexStu.gph qqstd.gph
```



Question Do the plots look exactly the same, similar or different (ignoring the scale of the y-axis)? Why?

The plots look similar, but NOT exactly the same, because the residuals are not multiplied by a constant multiple. Instead, the leverage for every point is different. The plots for the standardized and unadjusted residuals should look exactly the same (once we ignore y-axis scaling), because the unadjusted and

[3] Please note the Stata's definition for "standardized" residual is actually the "internally studentized" residual.

standardized residuals only differ by a constant multiple, which does not change their distributions.

Question What do these plots tell us?

Residuals are pretty nearly normally distributed, although there may be some outliers.

Leverage and Influence Diagnostics

To compute leverage, DFFITs and Cook's Distance for all observations, we can use the following Stata code.

```
* generate a new variable named 'h', that contains leverage values
predict h, leverage
* generate a new variable named 'cook', that contains cook's distance values
predict cook, cooksd
* generate a new variable named 'dfit', that contains dfitscvalues
predict dfit, dfits
```

Question What are some plots we can use to summarize these influence measures?

Individual histograms of each metric, leverage vs. studentized residuals, each metric plotted by observation number, etc.

Let's isolate some data points which violate our "rule of thumb" cut-offs for each diagnostic:

```
. clist person h sbp age smk quet if h> (2*(3+1))/34, noobs

  person          h       sbp       age       smk       quet
      10   .2645894       180        64         1      4.637
      33   .2933437       182        75         0        3.9
      34   .2886395        80        30         1        2.8

. clist person cook sbp age smk quet if cook>4/(34-2), noobs

  person       cook       sbp       age       smk       quet
       8   .1831151       160        48         1      3.612
       9   .2999472       144        44         1      2.368
      34   .9169716        80        30         1        2.8


. clist person dfit sbp age smk quet if abs(dfit)>2*sqrt((3+1)/34), noobs

  person       dfit       sbp       age       smk       quet
       8    .949097       160        48         1      3.612
       9   1.163967       144        44         1      2.368
      34   -2.25272        80        30         1        2.8
```

Question How many high leverage points are also influential? Is the outlier we identified visually before (#34) influential?

One point (#34) has high leverage and is influential according to both Cook's Distance and DFITS. Points 10 and 33 are high leverage by our rule of thumb, but are not influential according to their Cook's

Distance and DFITS. Points 8 and 9 are influential by both Cook's Distance and DFITS but do not reach the threshold of high leverage.

Refitting

Choose outlier(s) with unsatisfactory diagnostics and remove them from the data set before refitting the regression. We will choose to remove points 8, 9 and 34.

```
. drop if person==8 | person==9 | person==34
(3 observations deleted)

. regress sbp smk quet age
```

Source	SS	df	MS	Number of obs	=	31
Model	6559.99702	3	2186.66567	F(3, 27)	=	58.23
Residual	1013.93847	27	37.5532766	Prob > F	=	0.0000
Total	7573.93548	30	252.464516	R-squared	=	0.8661
				Adj R-squared	=	0.8513
				Root MSE	=	6.1281

sbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
smk	7.801385	2.246032	3.47	0.002	3.192907	12.40986
quet	7.498712	3.817304	1.96	0.060	-.3337479	15.33117
age	1.526173	.2331152	6.55	0.000	1.04786	2.004485
_cons	32.29686	8.84864	3.65	0.001	14.14096	50.45277

Question Did the removal of the outlier(s) substantially change our regression plot? When might this be a good thing? When might this be a bad thing? Would you drop the outlier(s)?

Yes, the coefficient for quet and the intercept changed dramatically.

When you remove an outlier, you're really saying that this patient/data point is in some way different from other patients/data points and thus shouldn't be used to analyze a relationship. The problem is, if you can't identify **why** this data point is different, you might end up overgeneralizing the relationship you're interested in.

For example, you're looking at how a drug affects blood pressure and most patients taking the drug result in lowered blood pressure compared to placebo. However, it is observed that a few patients result in much higher blood pressure on the treatment. If you don't take out these data points, the regression line which results may lead you to conclude that the drug does not work for anybody, whereas it really does work for a majority of people. If you just throw out these points, you may come to the conclusion that the drug does work for everybody. The truth may be, however, that these patients in the cluster of points shared a feature (for, example, pregnancy) which caused them to react badly to it, thus overgeneralizing the efficacy of the drug may be dangerous for this subset of patients.

In summary, outlier detection is an art and a science and must be treated thoughtfully.