Influential Observations and Regression Diagnostics

```
library(car)
library(ggplot2)
set.seed(1234)
```

Goals

- To distinguish among regression outliers, high-leverage observations, and influential observations and how to detect them.
- To introduce graphical methods for detecting non-normality, non-constant error variance, and nonlinearity.

Introduction

Linear statistical models make strong assumptions about the structure of data, which often do not hold in applications.

The method of least-squares is very sensitive to the structure of the data, and can be markedly influenced by one or a few unusual observations.

We could abandon linear models and least-squares estimation in favor of nonparametric regression and robust estimation. Alternatively, we can adapt and extend methods for examining and transforming data to diagnose problems with a linear model and to suggest solutions.

Outliers, Leverage, and Influence

Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis and because their presence may be a signal that the model fails to capture important characteristics of the data.

In simple regression, an outlier is an observation whose response variable value is conditionally unusual given the value of the explanatory variable.

In contrast, a univariate outlier is a value of Y or X that is unconditionally unusual; such a value may or may not be a regression outlier.

These distinctions are illustrated below for the simple regression model $Y = \beta_0 + \beta_1 X + \epsilon$

Regression outliers appear in (a) and (b).

In (a), the outlying observation has an X-value that is at the center of the X distribution; deleting the outlier has little impact on the least-squares fit.

In (b), the outlier has an unusual X-value; its deletion markedly affects both the slope and the intercept. Because of its unusual X- value, the outlying observation in (b) exerts strong leverage on the regression coefficients, while the outlying observation in (a) is a low-leverage point. The combination of high leverage with a regression outlier produces substantial influence on the regression coefficients.

In (c), the outlying observation has no influence on the regression coefficients even though it is a high-leverage point, because this observation is in line with the rest of the data.

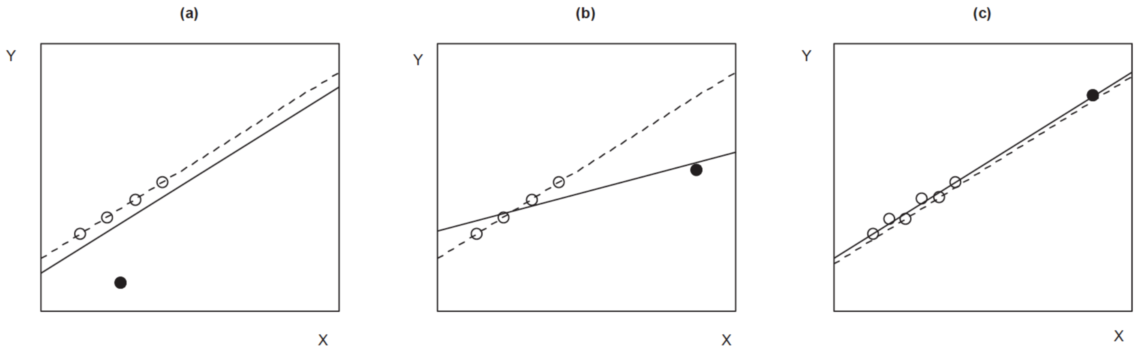


Figure 1: Unusual Observations

The following heuristic formula helps to distinguish among the three concepts of influence, leverage, and discrepancy ('outlyingness'):
Influence on Coefficients = Leverage × Discrepancy

Consider the example from the dataset with measured and reported weights of 183 male and female subjects who engage in programs of regular physical exercise.
This data could be modeled in two ways:

First, we could regress reported weight (RW) on measured weight (MW), a dummy variable for sex (F, coded 1 for women and 0 for men), and an interaction regressor (formed as the product $MW \times F$):

```
# The data are called Davis and are available in the car package
Davis$female <- ifelse(Davis$sex=="F",1,0)
mod1 <- lm(repwt ~ weight*female, data=Davis)
summary(mod1)
```

```
##
## Call:
## lm(formula = repwt ~ weight * female, data = Davis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.2230  -2.3247  -0.1325   2.0741  15.5783
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.35864    3.27719   0.415    0.679
## weight         0.98982    0.04260  23.236   <2e-16 ***
## female        39.96412    3.92932  10.171   <2e-16 ***
## weight:female -0.72536    0.05598 -12.957   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.661 on 179 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.8874, Adjusted R-squared:  0.8856
## F-statistic: 470.4 on 3 and 179 DF,  p-value: < 2.2e-16
```

$$\widehat{RW} = 1.36 + .99MW + 40F - .725(MW \times F)R^2 = 0.89 \text{ and } S_\epsilon = 4.66$$

Were these results taken seriously, we would conclude that men are unbiased reporters of their weights but women tend to over-report their weights if they are relatively light and under-report if they are relatively heavy.

The figure makes it clear that the differential results for women and men are due to one erroneous data point.

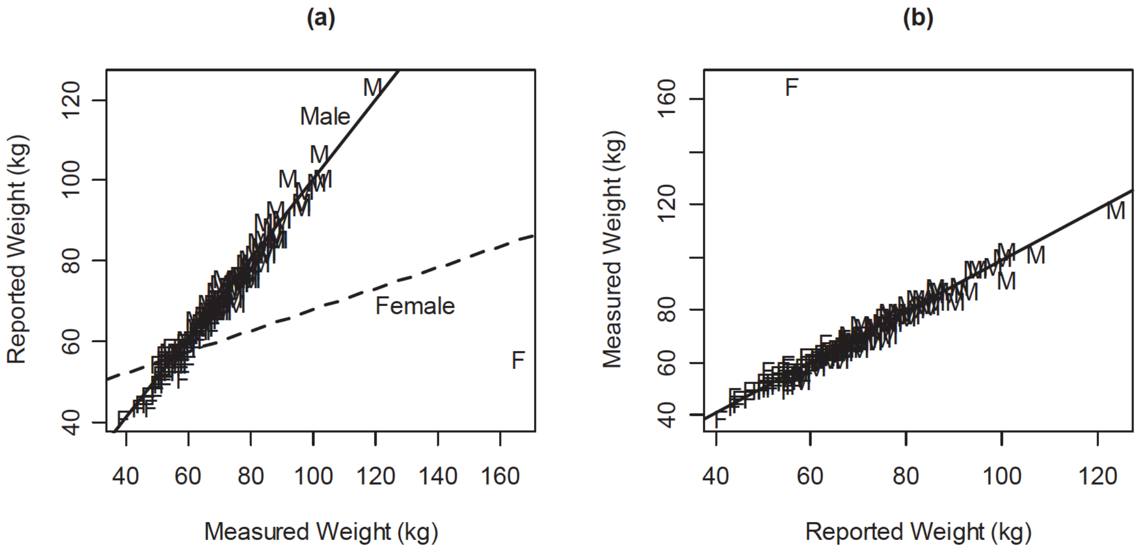


Figure 2: Regression of reported weight on measured weight, sex, and their interaction

Correcting the data produces the regression

```
cDavis <- Davis
cDavis[12, c(2, 3)] <- Davis[12, c(3, 2)]  # correct the recording error
mod2 <- lm(repwt ~ weight*female, data=cDavis)
summary(mod2)
```

```
##
## Call:
## lm(formula = repwt ~ weight * female, data = cDavis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4632 -1.1275  0.0717  1.2266  8.5777
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.35864    1.57725   0.861    0.390
## weight         0.98982    0.02050  48.279   <2e-16 ***
## female         1.98252    2.45028   0.809    0.420
## weight:female -0.05668    0.03845  -1.474    0.142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 179 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.9739, Adjusted R-squared:  0.9735
## F-statistic:  2229 on 3 and 179 DF,  p-value: < 2.2e-16
```

$$\widehat{RW} = 1.36 + .99MW + 1.98F - .0567(MW \times F)R^2 = 0.97 \text{ and } S_\epsilon = 2.24$$

Alternatively we could regress measured weight on reported weight, sex, and their interaction in the original data:

```
mod3 <- lm(weight ~ repwt*female, data=Davis)
summary(mod3)
```

```
##
## Call:
## lm(formula = weight ~ repwt * female, data = Davis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -7.655  -1.822  -0.757   0.655 108.405
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.794280   5.923944   0.303    0.762
## repwt         0.968918   0.076410  12.681   <2e-16 ***
## female        2.074211   9.297269   0.223    0.824
## repwt:female -0.009525   0.146855  -0.065    0.948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.449 on 179 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.6998, Adjusted R-squared:  0.6947
## F-statistic: 139.1 on 3 and 179 DF,  p-value: < 2.2e-16
```

$$\widehat{MW} = 1.79 + .969RW + 2.07F - .00953(RW \times F)R^2 = 0.70 \text{ and } S_\epsilon = 8.45$$

and in the corrected data:

```
mod4 <- lm(weight ~ repwt*female, data=cDavis)
summary(mod4)
```

```
##
## Call:
## lm(formula = weight ~ repwt * female, data = cDavis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6550 -1.0875 -0.2456  1.2833  6.3841
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.794280   1.579894   1.136    0.258
## repwt         0.968918   0.020378  47.547   <2e-16 ***
## female       -0.016776   2.479548  -0.007    0.995
## repwt:female  0.008306   0.039166   0.212    0.832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.253 on 179 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.9721, Adjusted R-squared:  0.9717
## F-statistic:  2082 on 3 and 179 DF,  p-value: < 2.2e-16
```

The outlier does not have much impact on the regression coefficients because the value of RW for the outlying observation is near $\overline{RW}$ for women.

There is, however, a marked effect on R^2 and the residual standard error: For the corrected data, $R^2 = 0.97$ and $S_\epsilon = 2.25$

Assessing Leverage

Hat Values

The hat-value, h_i is a common measure of leverage in regression. These values are so named because it is possible to express the fitted values, $\widehat{Y}_j$, in terms of the observed values, Y_i

$$\widehat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \ldots + h_{jj}Y_j + \ldots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i$$

Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value, $\widehat{Y}_j$: If h_{ij} is large, then the ith observation can have a substantial impact on the jth fitted value.

Properties of the hat-values:

$h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, and so the hat-value $h_i \equiv h_{ii}$ summarizes the potential leverage of Y_i on all of the fitted values

$1/n \le h_i \le 1$

The average hat-value is $\overline{h} = (k+1)/n$

In simple-regression analysis, the hat-values measure distance from the mean of X

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

In multiple regression, h_i measures distance from the centroid (point of means) of the X's, taking into account the correlational and variational structure of the X's, as illustrated for $k = 2$ below. Multivariate outliers in the X-space are thus high-leverage observations. The response-variable values are not at all involved in determining leverage.

For Davis's regression of reported weight on measured weight, the largest hat-value by far belongs to the 12th subject, whose measured weight was wrongly recorded as 166 kg.: $h_{12} = 0.714$. This quantity is many times the average hat-value, $\overline{h} = (3+1)/183 = 0.0219$.

Leverage can be obtained using the function `hatvalues(modelObjectName)`.

```
Davis$female <- ifelse(Davis$sex=="F",1,0)
mod1 <- lm(repwt ~ weight*female, data=Davis)
summary(hatvalues(mod1))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.009907 0.011133 0.013143 0.021858 0.018962 0.714186
```

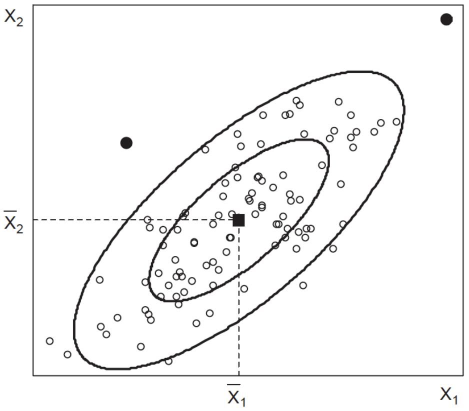


Figure 3: Contours of constant leverage in multiple regression with two explanatory variables. The two observations marked with solid black dots have equal hat-values.

Belsley, Kuh, and Welsch suggest that hat-values exceeding about twice the average $\overline{h} = (k+1)/n$ are noteworthy. In small samples, using $2 \times \overline{h}$ tends to nominate too many points for examination, and $3 \times \overline{h}$ can be used instead.

Detecting Outliers

Studentized Residuals

Discrepant observations usually have large residuals, but even if the errors ϵ_i have equal variances (as assumed in the general linear model), the residuals $\widehat{\epsilon}_i$ do not:

$$V(\widehat{\epsilon}_i) = \sigma_\epsilon^2(1 - h_i)$$

High-leverage observations tend to have small residuals, because these observations can coerce the regression surface to be close to them.

Residuals can be obtained using the function `resid(modelObjectName)`

Although we can form a standardized residual by calculating

$$\epsilon_i' = \frac{\epsilon_i}{S_\epsilon \sqrt{1 - h_i}}$$

this measure is slightly inconvenient because its numerator and denominator are not independent, preventing ϵ_i' from following a t-distribution: When $|\epsilon_i|$ is large, $S_\epsilon = \sqrt{\sum \epsilon_i^2/(n-k-1)}$, which contains ϵ_i^2, tends to be large as well.

Standardized residuals can be obtained using the function `rstandard(modelObjectName)`.

Suppose that we refit the model deleting the ith observation, obtaining an estimate $S_{\epsilon(-i)}$ of σ_ϵ that is based on the remaining $n-1$ observations.

Then the studentized residual

$$\epsilon_i^* = \frac{\epsilon_i}{S_{\epsilon(-i)} \sqrt{1 - h_i}}$$

has independent numerator and denominator, and follows a t-distribution with $n - k - 2$ degrees of freedom.

Studentized residuals can be obtained using the function `rstudent(modelObjectName)`.

Test for Outliers In most applications we want to look for any outliers that may occur in the data; we can in effect refit the model n times, producing studentized residuals $\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_n^*$. (It is not necessary to literally perform n auxiliary regressions.)

Usually, our interest then focuses on the largest absolute ϵ_i^*, denoted ϵ_{max}^*.

Because we have picked the largest of n test statistics, it is not legitimate simply to use t_{n-k-2} to find a p-value for ϵ_{max}^*.

One solution to this problem of simultaneous inference is to perform a Bonferroni adjustment to the p-value for the largest absolute ϵ_i^*:

Let $p' = Pr(t_{n-k-2} > \epsilon_{max}^*)$

Then the Bonferroni p-value for testing the statistical significance of ϵ_{max}^* is $p = 2np'$.

Note that a much larger ϵ_{max}^* is required for a statistically significant result than would be the case for an ordinary individual t-test.

In Davis's regression of reported weight on measured weight, the largest studentized residual by far belongs to the incorrectly coded 12th observation, with $\epsilon_{12}^* = -24.3$.

```
Davis$female <- ifelse(Davis$sex=="F",1,0)
mod1 <- lm(repwt ~ weight*female, data=Davis)
summary(rstudent(mod1))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -24.30446  -0.50377  -0.02849  -0.09618   0.44625   3.49663
```

Here, $n - k - 2 = 183 - 3 - 2$, and $Pr(t_{178} > 24.3) \approx 10^{-58}$.

The Bonferroni p-value for the outlier test is $p \approx 2 \times 183 \times 10^{-58} = 4 \times 10^{-56}$

```
outlierTest(mod1)  # function from the car package
```

```
##     rstudent unadjusted p-value Bonferroni p
## 12 -24.30446         1.9376e-58   3.5458e-56
```

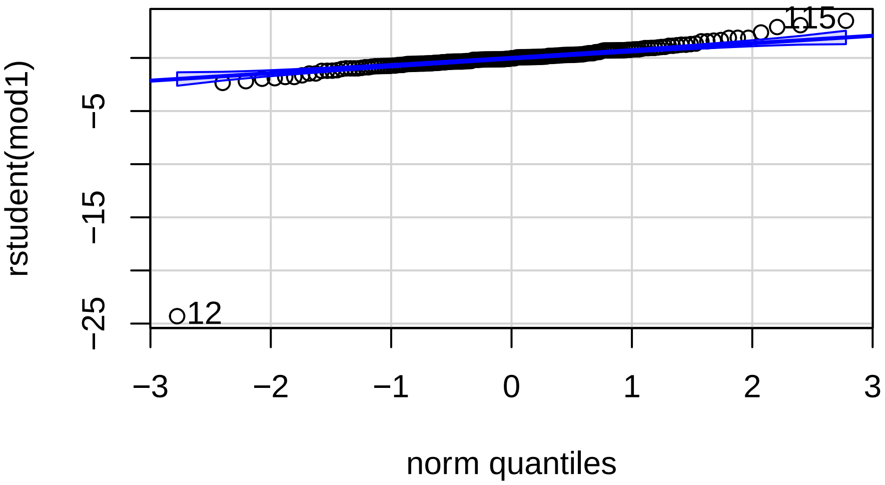
Beyond the issue of 'statistical significance,' it sometimes helps to call attention to residuals that are relatively large.

Under ideal conditions, about five percent of studentized residuals are outside the range $|\epsilon_i^* \leq 2|$. It is therefore reasonable to draw attention to observations outside this range.

For studentized residuals, outlier-testing provides a numerical cutoff, but this is no substitute for graphical examination of the residuals.

A graphical approach to detecting outliers is to construct a QQ plot for the studentized residuals, plotting against either the t or normal distribution.

```
qqPlot(rstudent(mod1)) # function from the car package
```



```
##  12 115
##  12 110
```

Measuring Influence

Influence on the regression coefficients combines leverage and discrepancy.

DFBETAS

The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

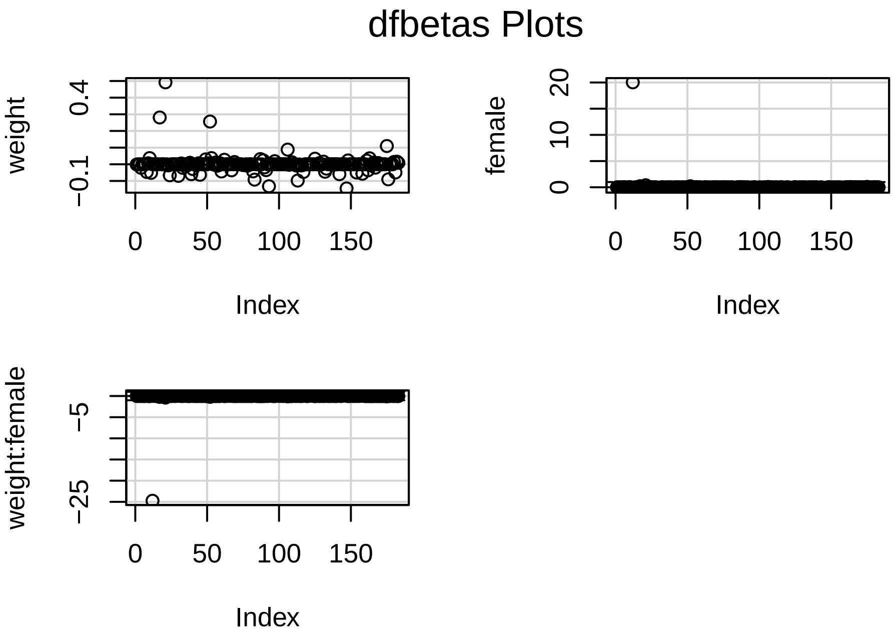
$$D_{ij} = \beta_j - \beta_{j(-i)} \text{ for } i = 1, \ldots, n \text{ and } j = 0, 1, \ldots, k$$

where the β_j are the least-squares coefficients calculated for all of the data, and the $\beta_{j(-i)}$ are the least-squares coefficients calculated with the ith observation omitted. These are called *DFBETAS* and are denoted as D_{ij}. Large values of *DFBETAS* indicate observations that are influential in estimating a given parameter.

Belsley, Kuh, and Welsch recommended $2/\sqrt{n}$ as a general cutoff value for DFBETAS to indicate influential observations, which would be 0.1478443 for the Davis data. However, another cut-off is to look for values > 1 or < -1.

DFBETAS can be obtained using the function `dfbetas(modelObjectName)`. The `dfbetasPlots(modelObjectName)` function from the `car` package will plot the DFBETAS.

```
Davis$female <- ifelse(Davis$sex=="F",1,0)
mod1 <- lm(repwt ~ weight*female, data=Davis)
dfbetasPlots(mod1)
```



One problem associated with using the D_{ij} is there are a large number of them, specifically $n(k+1)$. It is useful to have a single summary index of the influence of each observation on the least-squares fit.

Cook's Distance

Cook (1977) proposed measuring the 'distance' between the β_j and the corresponding $\beta_{j(-i)}$ by calculating the F-statistic for the 'hypothesis' that $\beta_j = \beta_{j(-1)}$, for $j = 0, 1, \ldots, k$

This statistic is recalculated for each observation, $i = 1, \ldots, n$

The resulting values should not literally be interpreted as F-tests, but rather as a distance measure that does not depend upon the scales of the X's.

Cook's statistic can be written (and simply calculated) as

$$D_i = \frac{\epsilon_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

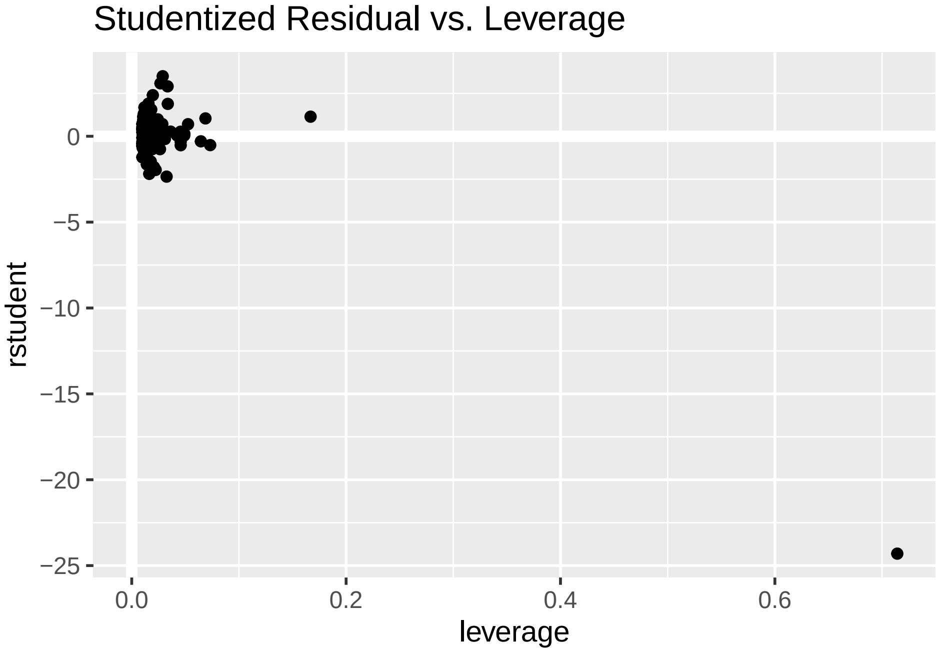
In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second is a measure of leverage.

We look for values of D_i that are substantially larger than the rest.

Because all of the deletion statistics depend on the hat-values and residuals, a graphical alternative is to plot ϵ_i^* against h_i and look for observations for which both are large.

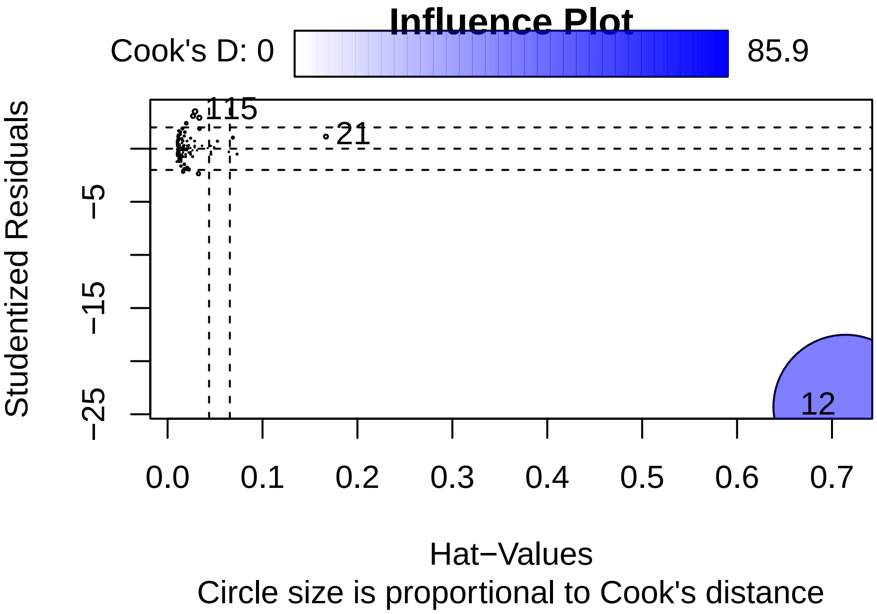
```
rstudent <- rstudent(mod1)
leverage <- hatvalues(mod1)
ggplot( , aes(leverage, rstudent)) +
  geom_vline(size = 2, colour = "white", xintercept = 0) +
  geom_hline(size = 2, colour = "white", yintercept = 0) +
  geom_point() +
  ggtitle("Studentized Residual vs. Leverage")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



A slightly more sophisticated version of this plot that incorporates Cook's D can be obtained using the `influencePlot` function from the `car` package.

```
influencePlot(mod1, main="Influence Plot",
              sub="Circle size is proportional to Cook's distance")
```



```
##        StudRes        Hat       CookD
## 12  -24.304463 0.71418565 85.92734587
## 21    1.141619 0.16684054  0.06513595
## 115   3.496628 0.02891086  0.08562939
```

For Davis's regression of reported weight on measured weight, Cook's D points to the obviously discrepant

12th observation:

Cook's $D_{12} = 85.9$ (next largest, $D_{21} = 0.065$)

The following size-adjusted cutoff for Cook's D was suggested by Chatterjee and Hadi:

$$D_i > \frac{4}{n-k-1}$$

Absolute cutoffs for D, such as $D_i > 1$, risk missing relatively influential data.

A line can be drawn on a graph at the value of a numerical cutoff, and observations that exceed the cutoff can be identified individually.

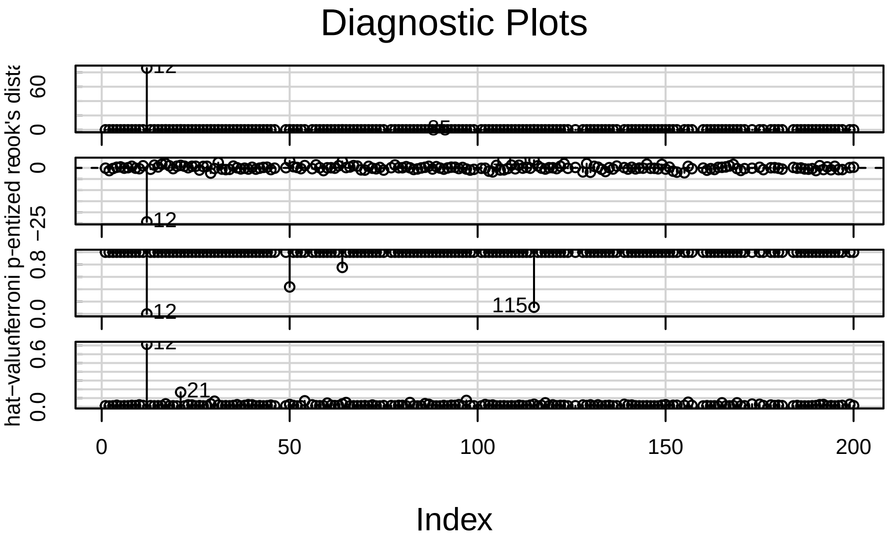
Cook's Distance can be obtained using the function `cooks.distance(modelObjectName)`.

```
cooksD <- cooks.distance(mod1)
n <- 183
k <- 3
cutoff <- 4/(n-k-1)
cutoff
```

```
## [1] 0.02234637
```

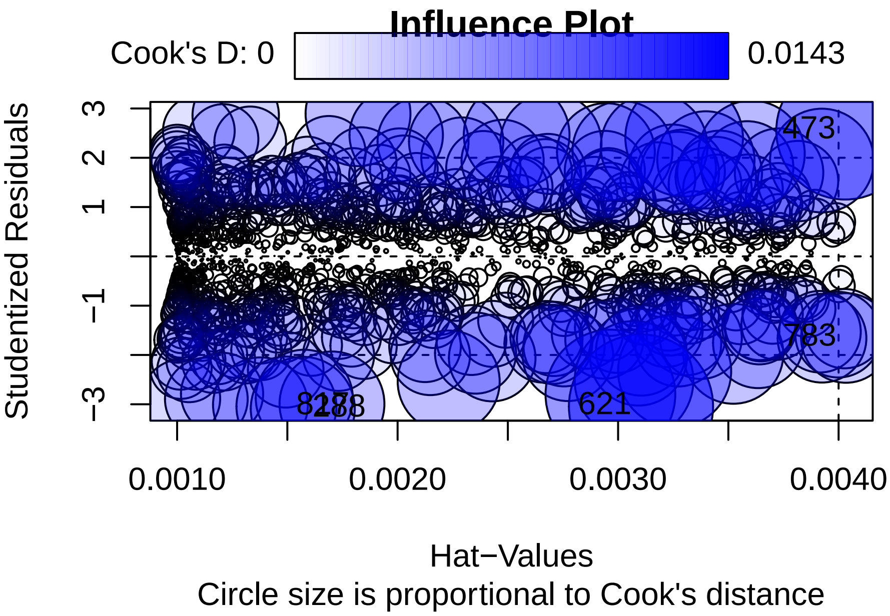
Graph of Cook's D by observation number:

```
infIndexPlot(mod1)
```



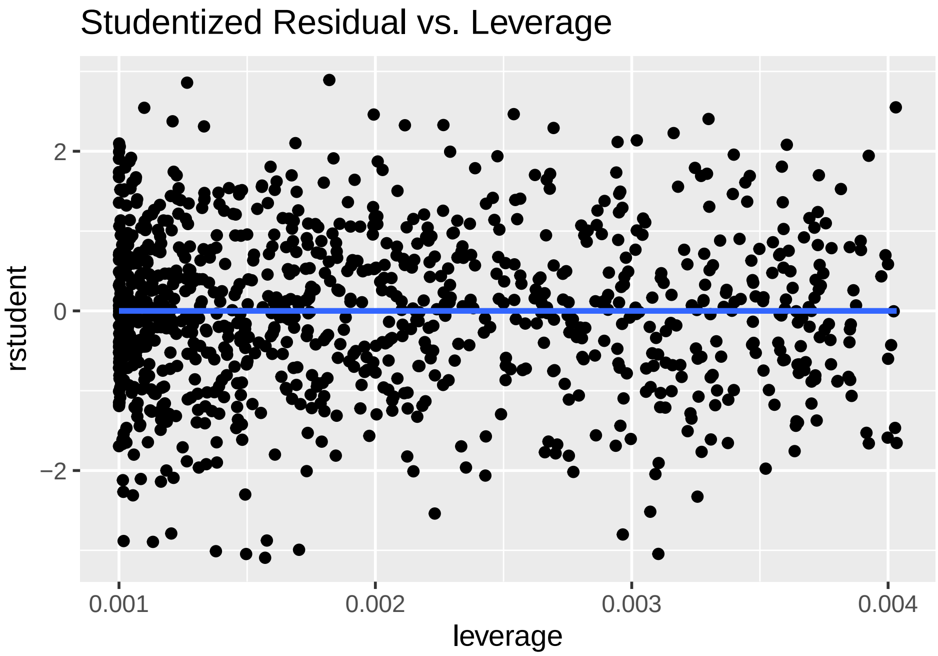
An ideal case

The plot below is a great example of a Studentized Residuals vs Leverage plot in which we see no evidence of outliers. None of the points come close to having both high residual and leverage.

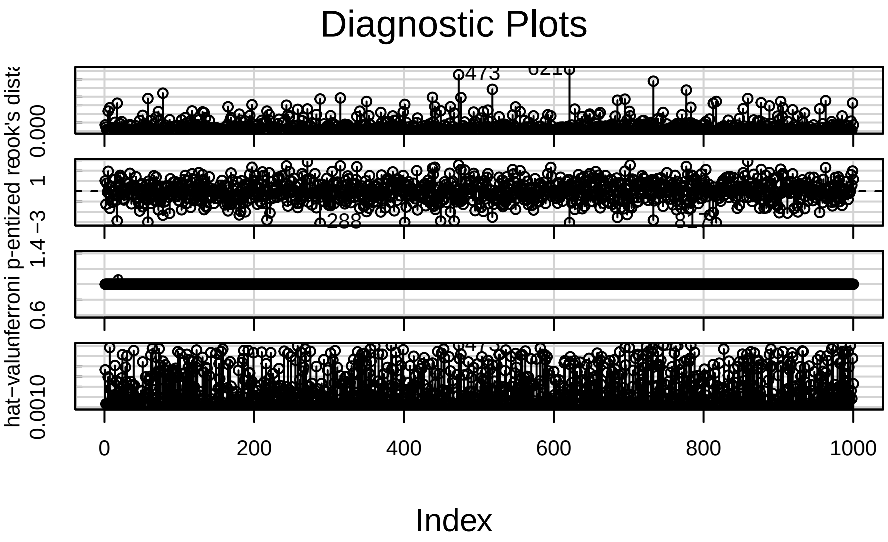


```
##        StudRes         Hat       CookD
## 288 -3.093109 0.001569940 0.007457847
## 473  2.550221 0.004030140 0.013086130
## 621 -3.044920 0.003103869 0.014314976
## 783 -1.655168 0.004033375 0.005537600
## 817 -3.046056 0.001495918 0.006893121
```

Without the circle size proportional to Cook's distance:



Graph of Cook's D by observation number:



DFFITS *DFFITS* can also be used to identify influential data points. They are a standardized function of the difference between the fitted value for the ith observation when it is included and when it is excluded (i.e., deleted) and the model re-fit. It quantifies the number of standard deviations that the fitted value changes when the ith data point is omitted.

An observation is deemed influential if the absolute value of its DFFITS value is greater than: $|DFFITS| > 2\frac{\sqrt{(k+1)}}{(n-k-1)}$

DFFITS can be obtained using the function `dffits(modelObjectName)`.

We have focused on changes in regression coefficients but other regression outputs, such as the coefficient sampling variances and covariances, are also subject to influence.

Assumptions of the Linear Regression Model

- *Linearity* — If the dependent variable is linearly related to the independent variables, there should be no systematic relationship between the residuals and the predicted (i.e., fitted) values.
- *Normality* — If the dependent variable is normally distributed for a fixed set of predictor values, then the residual values should be normally distributed with a mean of 0. The Normal Q-Q plot is a probability plot of the standardized residuals against the values that would be expected under normality. If the normality assumption is met, the points on this graph should fall on a straight 45-degree line.
- *Homoscedasticity* — The variance of Y around the regression line is constant. If the constant variance assumption is met, the points in the Scale-Location graph should be a random band around a horizontal line.

There is another assumption that cannot be plotted, that of *independence*, which states that the Y_i values (and thus the residuals) are independent of each other. The best way to assess whether the dependent variable values are independent is from your knowledge of how the data were collected. If there are repeated measures on the same individual or if the data were from a sample of siblings, then this assumption is violated. For this type of data, you need to use multilevel regression models.

Plots are frequently made to assess these assumptions.

Nonlinearity

The linearity assumption is easily checked by plotting the studentized residuals vs. the fitted values. If linearity holds, then there is no systematic relationship. But if linearity does not hold, then it will be

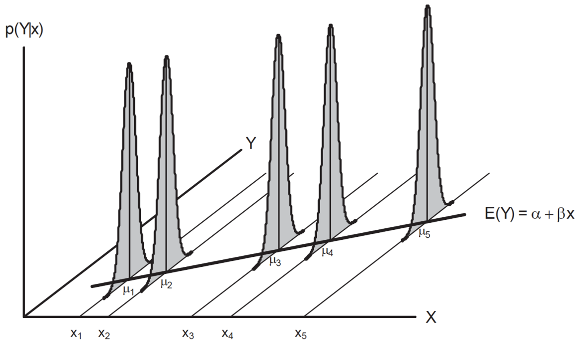
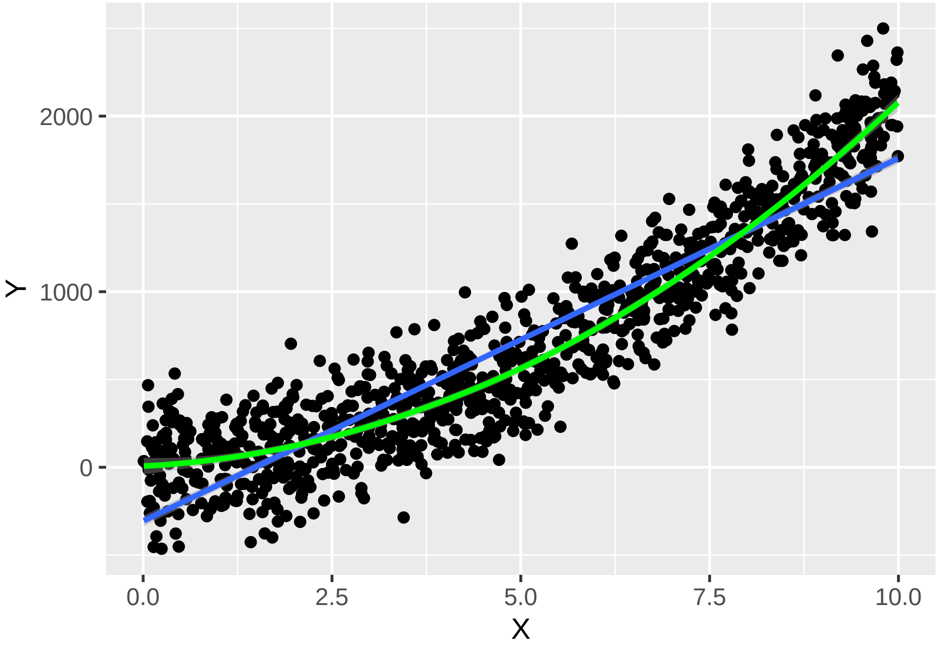


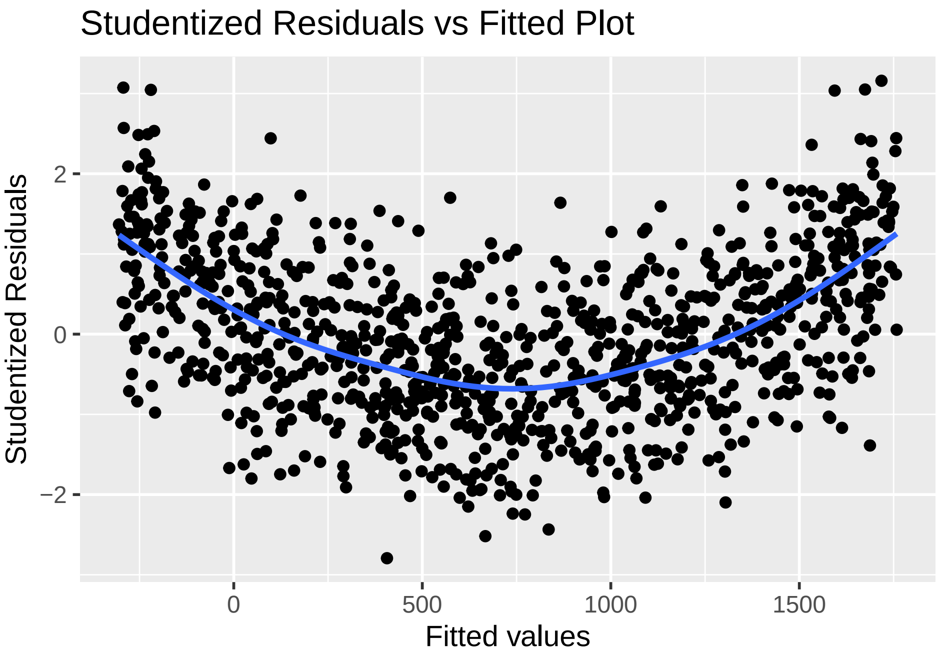
Figure 4: Usual Assumptions

apparent.

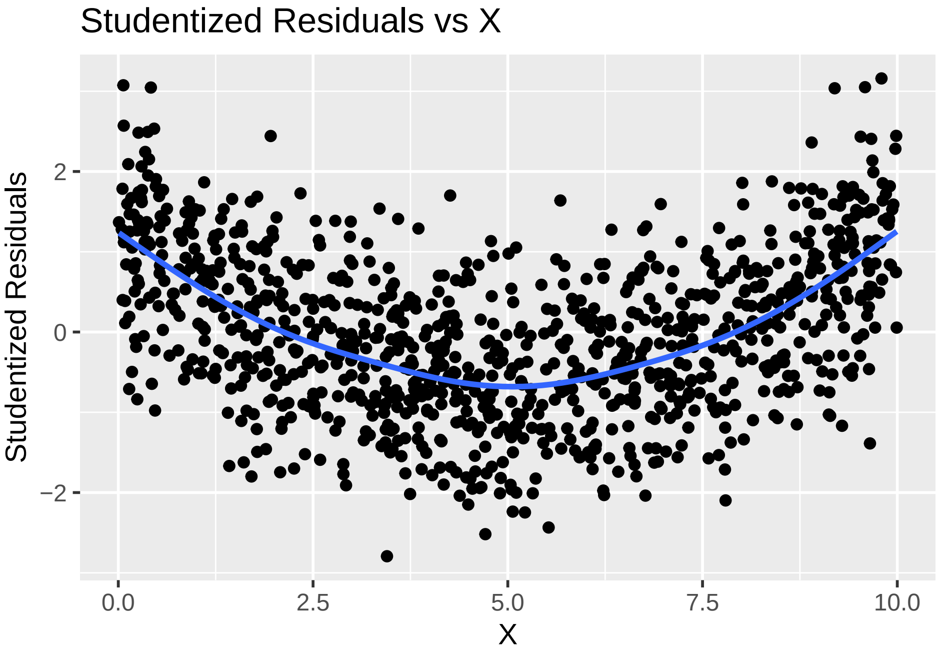
For example, here is a scatterplot of two variables, X and Y.



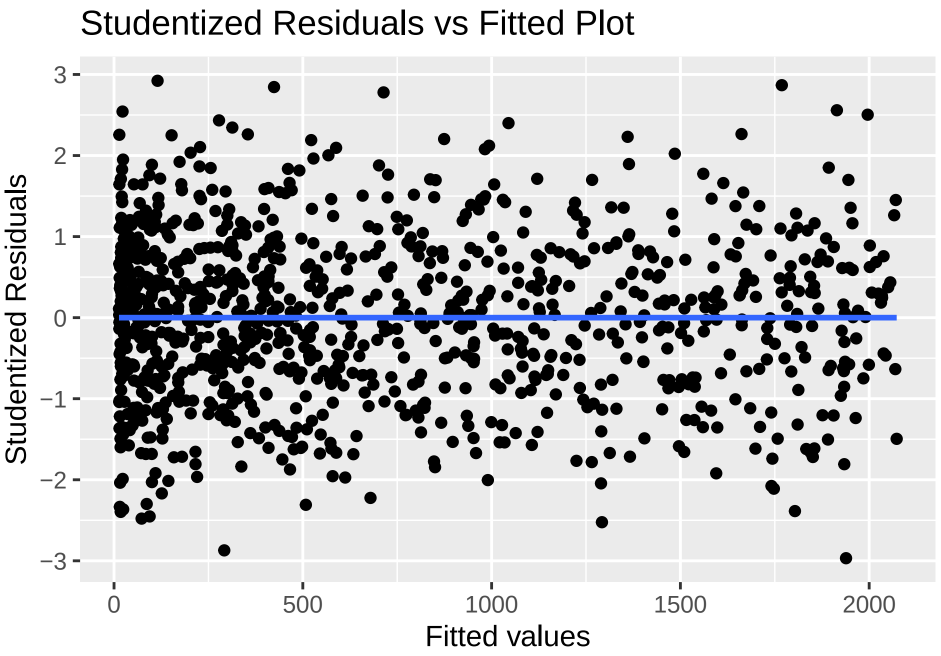
Suppose we fit a regression of Y on X and obtain the fitted values and studentized residuals. Here is a plot of the studentized residuals vs. the fitted values. It is clear that there is a systematic relationship.



Plotting studentized residuals against each X is frequently helpful for detecting departures from linearity.



The plot suggests a polynomial relationship so let's include a quadratic term in the regression and plot the studentized residuals and fitted values from this regression.



There is no longer a systematic relationship.

Added-Variable and Component-Residual Plots

Added Variable Plots: Joint Influence

Added-variable ('partial-regression') plots may be used to display leverage and influence on particular coefficients.

As illustrated below, subsets of observations can be jointly influential or can offset each other's influence.

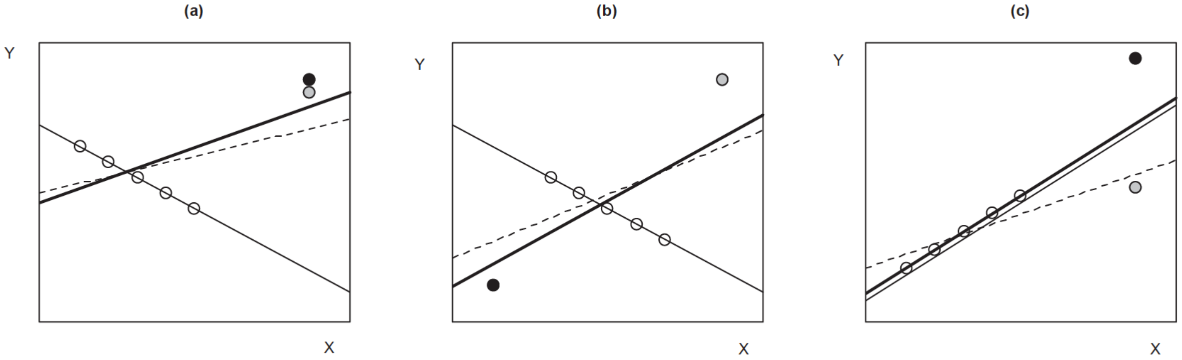


Figure 5: Jointly Influential Observations

The heavy solid line is the least-squares regression line for all the data. The broken line deletes the black solid point and the lighter solid line deletes both the black and grey solid points.

In a) there is a pair of jointly influential points.

In b) the jointly influential pair is widely separated.

In c) the two points offset each other.

Influential subsets or multiple outliers can often be identified by applying single-observation diagnostics, such as Cook's D and studentized residuals, sequentially.

It can be important to refit the model after deleting each point, because the presence of a single influential value can dramatically affect the fit at other points.

Although it is possible to generalize deletion statistics to subsets of several points, the very large number of subsets usually renders this approach computationally impractical.

An attractive alternative is to employ graphical methods, and a particularly useful influence graph is the added-variable plot (also called a partial-regression plot or an partial-regression leverage plot) which can be obtained using the `avPlots(modObjectName)` function from the `car` package.

Added variable plots are constructed as follows:

Let $Y_i^{(1)}$ represent the residuals of the least squares regression of Y on all of the X's with the exception of X_1.

$$Y_i = \beta_0^{(1)} + \beta_2^{(1)} X_{i2} + \ldots + \beta_k^{(1)} X_{ik} + Y_i^{(1)}$$

Likewise, $X_i^{(1)}$ are the residuals from the least squares regression of X_1 on all the other X's.

$$X_{i1} = \gamma_0^{(1)} + \gamma_2^{(1)} X_{i2} + \ldots + \gamma_k^{(1)} X_{ik} + X_i^{(1)}$$

So $Y_i^{(1)}$ and $X_i^{(1)}$ are the parts of Y and X_1 that remain when the effects of $X_2, \ldots, X_k$ are removed.

$Y_i^{(1)}$ and $X_i^{(1)}$ have the following interesting properties:

1. The slope from the least squares regression of $Y^{(1)}$ and $X^{(1)}$ is the least squares slope, β_1 from the full model.
2. The residuals from the simple regression of $Y^{(1)}$ and $X^{(1)}$ are the same as those from the full regression.
3. The variation of $X^{(1)}$ is the conditional variation of X_1 holding the other X's constant and therefore the standard error of β_1 in the simple regression is the same as that from the full model (except for df).

Plotting $Y^{(1)}$ against $X^{(1)}$ allows examination of leverage and influence on β_1 and also gives a visual impression of the precision of estimation for β_1.

Similar added variable plots can be constructed for other regression coefficients: Plot $Y^{(j)}$ against $X^{(j)}$ for each $j = 0, \ldots, k$

Component Residual Plots: Detecting Nonlinearity

Although it is useful in multiple regression to plot Y against each X, these plots can be misleading, because our interest centers on the partial relationship between Y and each X, controlling for the other X's, not on the marginal relationship between Y and an individual X, ignoring the other X's.

Plotting studentized residuals against each X is frequently helpful for detecting departures from linearity.

However, these residual plots cannot distinguish between monotone and non-monotone nonlinearity.

The distinction is important because monotone nonlinearity frequently can be 'corrected' by simple transformations. Thus, it is necessary to focus on particular patterns of departure from linearity.

Case (a) might be modeled by $Y = \beta_0 + \beta_1 \sqrt{X} + \epsilon$.

Case (b) cannot be linearized by a power transformation of X, and might instead be dealt with by the quadratic regression, $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$.

Added-variable plots, introduced previously for detecting influential data, can reveal nonlinearity and suggest whether a relationship is monotone.

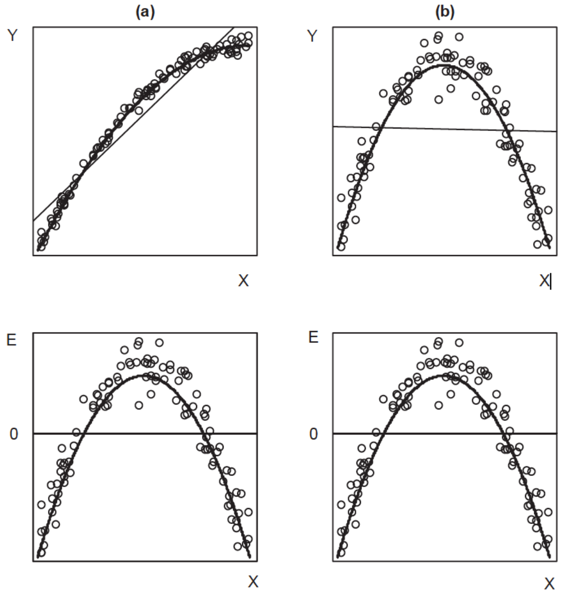


Figure 6: Residual plots are identical even though the regression of Y on X is monotone in a) and non-monotone in b).

However, these plots are not always useful for locating a transformation. The added-variable plot adjusts X_j for the other X's, but it is the unadjusted X_j that is transformed in respecifying the model.

Component+residual plots, also called partial-residual plots (as opposed to partial-regression = added-variable plots) are often an effective alternative.

Component+residual plots are not as suitable as added-variable plots for revealing leverage and influence.

The partial residual for the jth explanatory variable is

$$\epsilon_i^{(j)} = \epsilon_i + \beta_j X_{ij}$$

In words, add back the linear component of the partial relationship between Y and X_j to the least-squares residuals, which may include an unmodeled nonlinear component.

Then plot $\epsilon^{(j)}$ vs. X_j.

By construction, the multiple-regression coefficient β_j is the slope of the simple linear regression of $\epsilon^{(j)}$ on X_j, but nonlinearity may be apparent in the plot as well.

Normality Assumption

Normally distributed errors

The assumption of normally distributed errors is almost always arbitrary but the central-limit theorem assures that inference based on the least squares estimator is approximately valid. So then why should we be concerned about non-normally distributed errors?

Although the validity of least-squares estimation is robust, the efficiency of least squares is not: The least-squares estimator is maximally efficient among unbiased estimators when the errors are normally distributed. For 'heavy-tailed' errors, the efficiency of least-squares estimation decreases markedly.

Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit as a conditional typical value of Y.

A multimodal error distribution suggests the omission of one or more discrete explanatory variables that divide the data naturally into groups.

QQ plots are useful for examining the distribution of the residuals, which are estimates of the errors.

We compare the sample distribution of the studentized residuals, $\widehat{\epsilon}_i^*$, with the quantiles of the standard normal distribution, or with those of the t-distribution for $n - k - 2$ degrees of freedom.

Even if the model is correct, the studentized residuals are not an independent random sample from t_{n-k-2}. Correlations among the residuals depend upon the configuration of the X-values, but they are generally negligible unless the sample size is small. At the cost of some computation, it is possible to adjust for the dependencies among the residuals in interpreting a QQ plot.

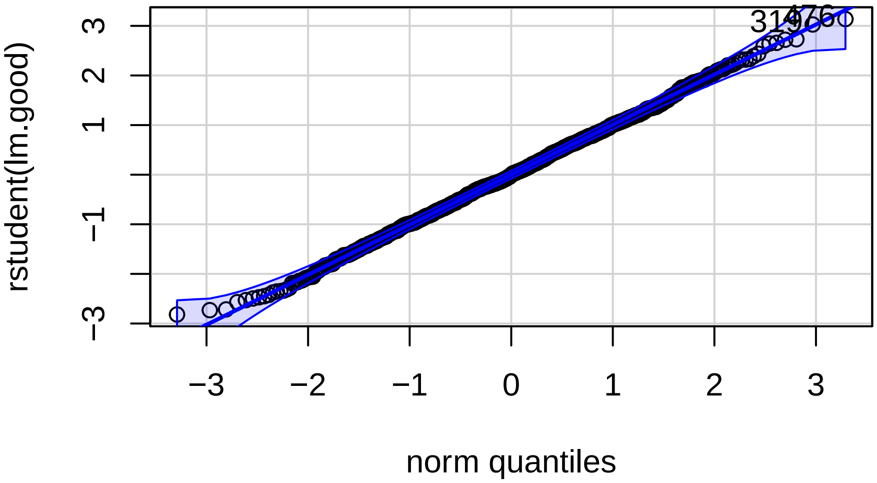
The QQ plot is effective in displaying the tail behavior of the residuals: Outliers, skewness, heavy tails, or light tails all show up clearly.

Other univariate graphical displays, such as histograms and density estimates, can supplement the QQ plot.

An ideal case Below I generate some data according to:

$$Y_i = 3 + 0.1X_i + \epsilon_i$$

for $i = 1, 2, \ldots, 1000$, where the ϵ_i are independent normal variables (with standard deviation 3), $N(0, sd = 3)$, which perfectly satisfies all the of the standard assumptions of linear regression.



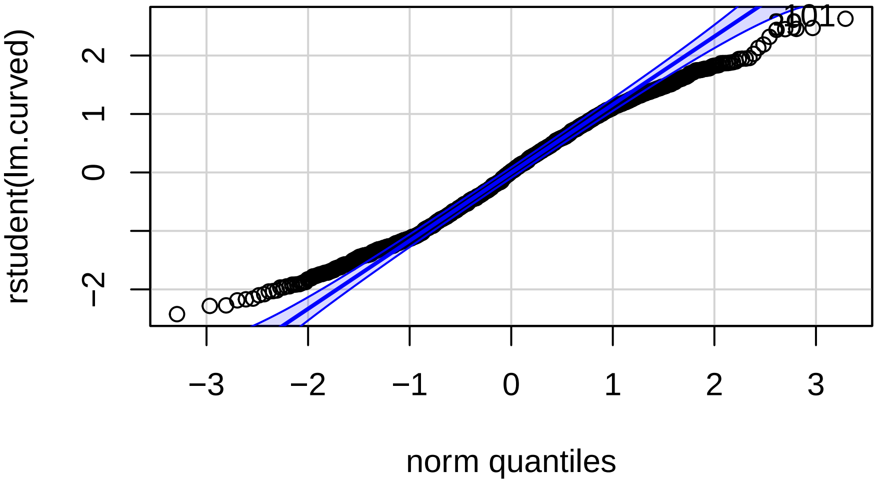
```
## [1] 476 319
```

The residuals here are a perfect match to the diagonal line, and thus are normally distributed. This is an example of a Normal QQ plot that is as perfect as it gets.

If we use the outlier test, we do not find any with a Bonferonni adjusted $p < 0.05$.

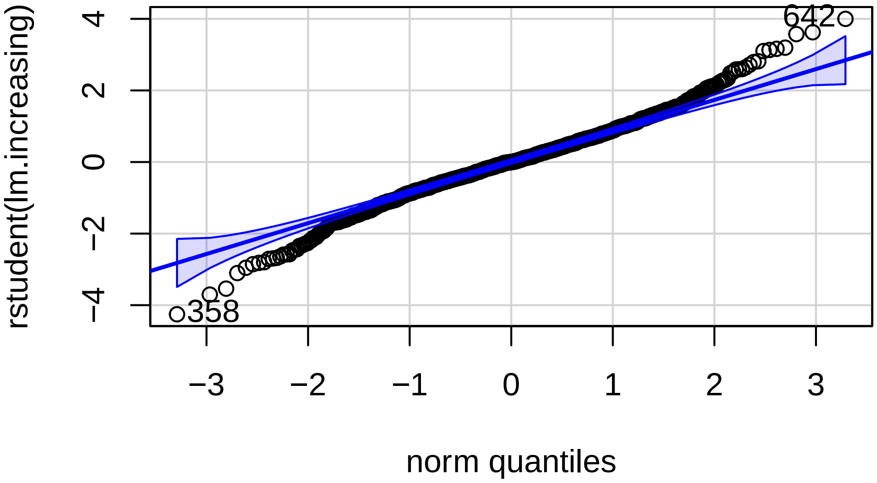
```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 476 3.136826          0.0017579           NA
```

Lighter Tails In the next example, we see a QQ plot where the residuals deviate from the diagonal line in both the upper and lower tail. This plot indicates that the tails are 'lighter' (have smaller values) than what we would expect under the standard modeling assumptions. This is indicated by the points forming a "flatter" line than than the diagonal.



```
## [1] 101  39
```

Heavier Tails In this final example, we see a QQ plot where the residuals again deviate from the diagonal line in both the upper and lower tail. Unlike the previous plot, in this case we see that the tails are observed to be 'heavier' (have larger values) than what we would expect under the standard modeling assumptions. This is indicated by the points forming a "steeper" line than the diagonal.



```
## [1] 358 642
```

Constant Error Variance

Although the least-squares estimator is unbiased and consistent even when the error variance is not constant, its efficiency is impaired, and the usual formulas for coefficient standard errors are inaccurate. Non-constant error variance is sometimes termed 'heteroscedasticity.'

Because the regression surface is k-dimensional, and imbedded in a space of $k + 1$ dimensions, it is generally impractical to assess the assumption of constant error variance by direct graphical examination of the data.

It is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X. The former situation can often be detected by plotting residuals against fitted values; the latter by plotting residuals against each X.

Plotting residuals against Y (as opposed to $\widehat{Y}$) is generally unsatisfactory, because the plot will be 'tilted' as there is a built-in linear correlation between Y and $\widehat{\epsilon}$, because $Y = \widehat{Y} + \epsilon$.

The least-squares fit insures that the correlation between $\widehat{Y}$ and $\widehat{\epsilon}$ is zero, producing a plot that is much easier to examine for evidence of non-constant spread.

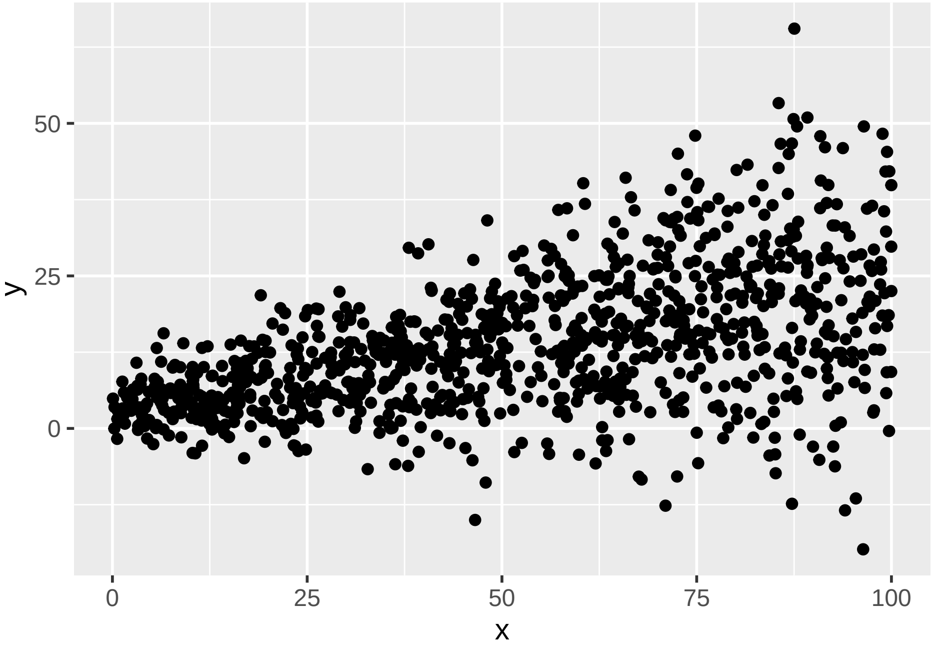
Because the residuals have unequal variances even when the variance of the errors is constant, it is preferable to plot studentized residuals against fitted values.

It often helps to plot $\sqrt{|\hat{\epsilon}_i^*|}$ against $\widehat{Y}$, which is referred to as a scale-location plot.

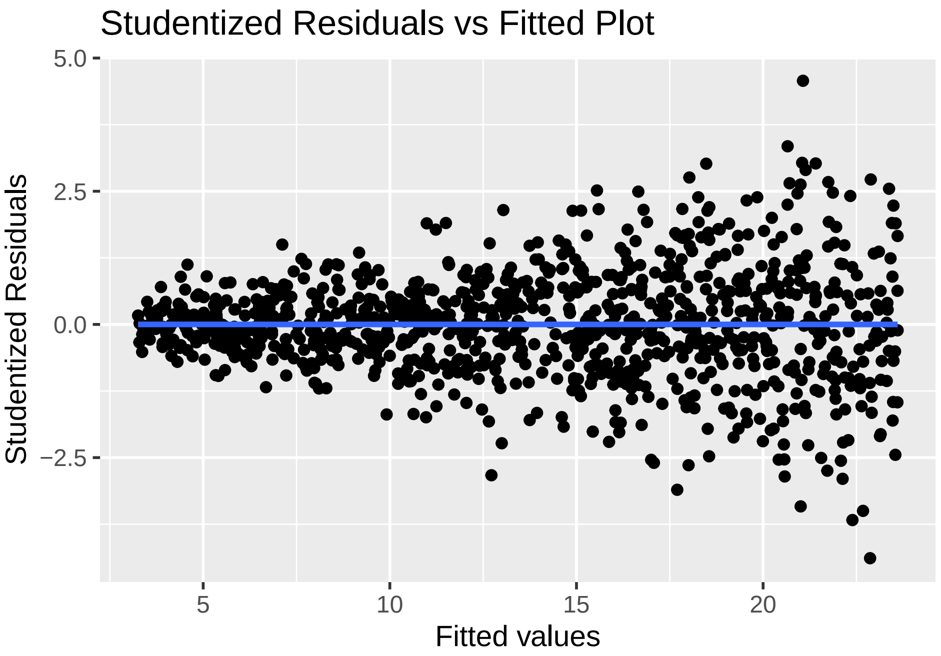
Non-constant error variance is a serious problem only when it is relatively extreme — say when the magnitude (i.e., the standard deviation) of the errors varies by more than a factor of about three — that is, when the largest error variance is more than about 10 times the smallest (although there are cases where this simple rule fails to offer sufficient protection).

Below I generate data in which the assumption does not hold.

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

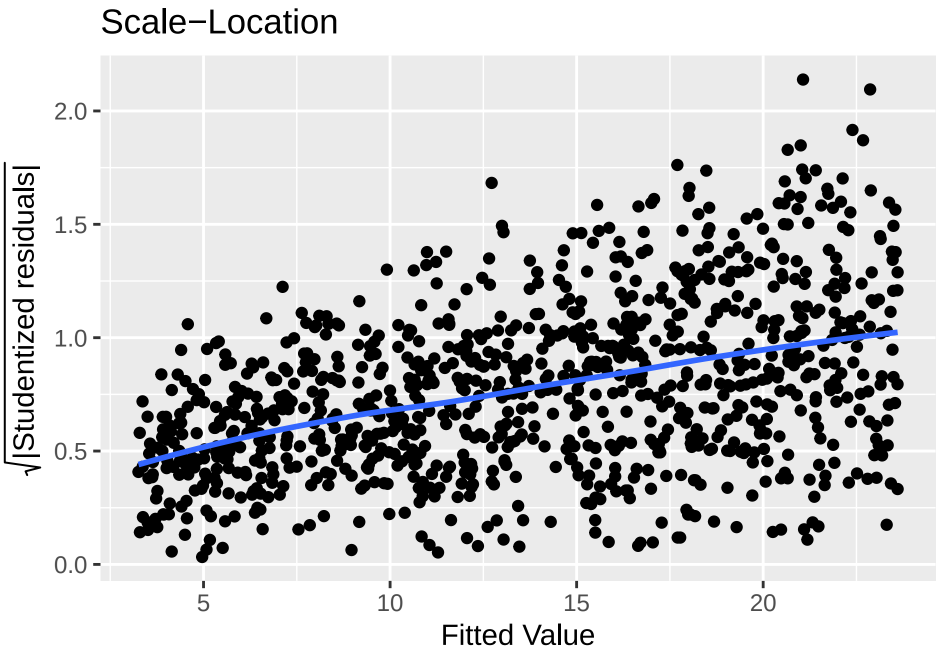


Here is what the Studentized Residuals vs. Fitted plot looks like in this case.



The distribution of the residuals is quite well concentrated around 0 for small fitted values, but the spread increases as the fitted values increase. This is an instance of "increasing variance". The standard linear regression assumption is that the variance is constant across the entire range. When this assumption is not valid, such as in this example, we should not believe our confidence intervals, prediction bands, or the p-values in our regression.

Scale-Location Plot



The `ncvTest()` function in the `car` package performs a statistical test of the null hypothesis of constant error variance against the alternative that the error variance changes with the level of the fitted values. A significant result suggests heteroscedasticity (nonconstant error variance).

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 231.2401, Df = 1, p = < 2.22e-16
```

Using this test, we find it is statistically significant. Thus, we would reject the null hypothesis that the error variance is constant.

Transformations

Introduction

'Classical' statistical models make strong assumptions about the structure of data, assumptions which often fail to hold in practice. One solution is to abandon classical methods. Another solution is to transform the data so that they conform more closely to the assumptions. Transformations can also often assist in the examination of data in the absence of a statistical model. A particularly useful group of transformations is the 'family' of powers and roots: $X \rightarrow X^p$

If p is negative, then the transformation is an inverse power: $X^{-1} = 1/X$ and $X^{-2} = 1/X^2$. If p is a fraction, then the transformation represents a root: $X^{1/2} = \sqrt{X}$ and $X^{-1/2} = 1/\sqrt{X}$

It is sometimes convenient to define the family of power transformations in a slightly more complex manner (called the Box-Cox family):

$$X \rightarrow X^{(p)} \equiv \frac{X^p - 1}{p}$$

Because $X^{(p)}$ is a linear function of X^p, the two transformations have the same essential effect on the data, but $X^{(p)}$ reveals the essential unity of the family of powers and roots.

Dividing by p preserves the direction of X, which otherwise would be reversed when p is negative.

The transformations $X^{(p)}$ are 'matched' above $X = 1$ both in level and slope.

The power transformation X^0 is useless, but the very useful log transformation is a kind of 'zeroth' power because

$$\lim_{p \rightarrow 0} \frac{X^p - 1}{p} = \log_e X$$

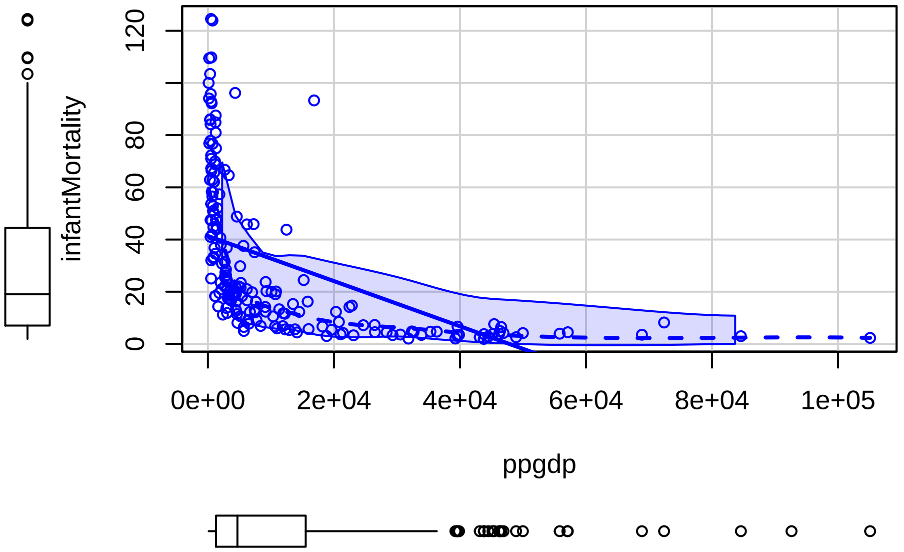
where $e \approx 2.718$ is the base of natural logarithms. It is generally more convenient to use logs to the base 10 or base 2, which are more easily interpreted than logs to the base e. Changing bases is equivalent to multiplying by a constant.

Power transformations are sensible only when all of the values of X are positive. First of all, some of the transformations, such as log and square root, are undefined for negative or zero values. Second, the power transformations are not monotone when there are both positive and negative values in the data. We can add a positive constant (called a 'start') to each data value to make all of the values positive: $X \rightarrow (X + s)^p$

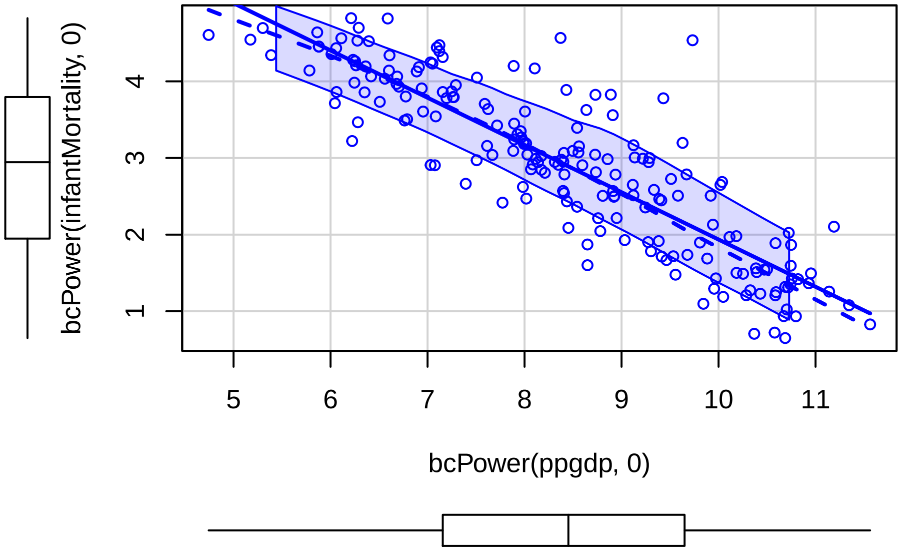
Power transformations are effective only when the ratio of the largest data values to the smallest ones is sufficiently large; if this ratio is close to 1, then power transformations are nearly linear.

Below is a scatterplot of infant mortality and income from the `UN` data that is included in the `car` package. The `scatterplot` function is included in the `car` package and by default it includes a regression line (the solid line), a loess smoother (the middle dashed line), and marginal boxplots on each axis. The `bcPower()` function, also contained in the `car` package, allows various Box-Cox power transformations.

```
# original scaling
scatterplot(infantMortality ~ ppgdp, data=UN)
```



```
# log transformations
scatterplot(bcPower(infantMortality, 0) ~ bcPower(ppgdp, 0), data=UN)
```



```
# code below does the same thing
# scatterplot(infantMortality ~ ppgdp, log="xy", data=UN)
```

Transforming skewness Highly skewed distributions are difficult to examine.

Apparently outlying values in the direction of the skew are brought in towards the main body of the data.

Unusual values in the direction opposite to the skew can be hidden prior to transforming the data.

Statistical methods such as least-squares regression summarize distributions using means. The mean of a skewed distribution is not a good summary of its center.

Power transformations can make a skewed distribution more symmetric.

Descending the ladder of powers to $\log X$ makes the distribution more symmetric by pulling in the right tail.

Ascending the ladder of powers to X^2 or X^3 can ‘correct’ a negative skew.

If we have a choice between transformations that perform roughly equally well, we may prefer one transformation to another because of interpretability. The log transformation has a convenient multiplicative interpretation (e.g., adding 1 to $\log_2 X$ doubles X, adding 1 to $\log_{10} X$ multiplies X by 10).

Transforming nonlinearity Power transformations can also be used to make many nonlinear relationships more nearly linear.

Linear relationships — expressible in the form $\widehat{Y} = a + bX$ are particularly simple. There is a simple and elegant statistical theory for linear models.

When there are several explanatory variables, the alternative of nonparametric regression may not be feasible or may be difficult to visualize.

There are certain technical advantages to having linear relationships among the explanatory variables in a regression analysis.

Mosteller and Tukey's 'bulging rule' can be used to select a transformation.

For example, if the 'bulge' points down and to the right, transform Y down the ladder of powers or X up the ladder of powers (or both).

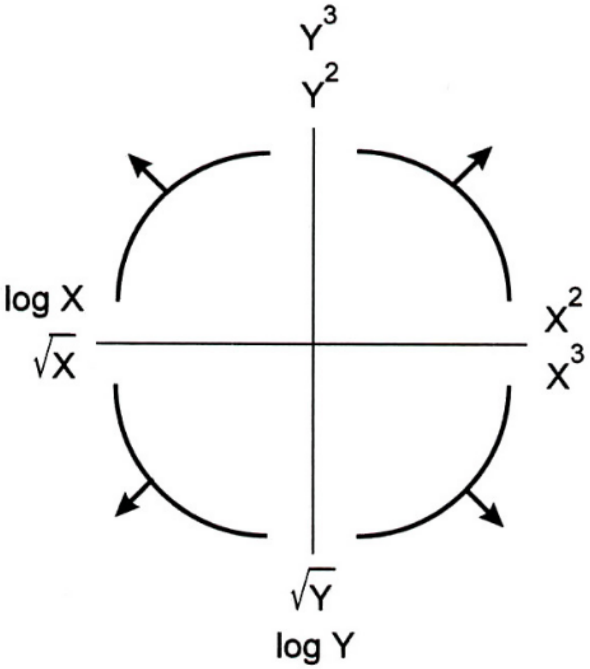


Figure 7: Bulging Rule

Since the bulge points up and to the left, we can try transforming prestige up the ladder of powers or income down. The cube-root transformation of income works reasonably well.

The bulging rule suggests that infant mortality or income or both should be transformed down the ladder of powers and roots because the bulge points down and to the left.

Transforming non-constant spread When a variable has very different degrees of variation in different groups, it becomes difficult to examine the data and to compare differences in level across the groups.

Differences in spread are often systematically related to differences in level.

Tukey suggested graphing the log hinge-spread against the log median. The slope of the linear 'trend,' if any, in the spread-level plot can be used to suggest a spread-stabilizing power transformation of the data. Express

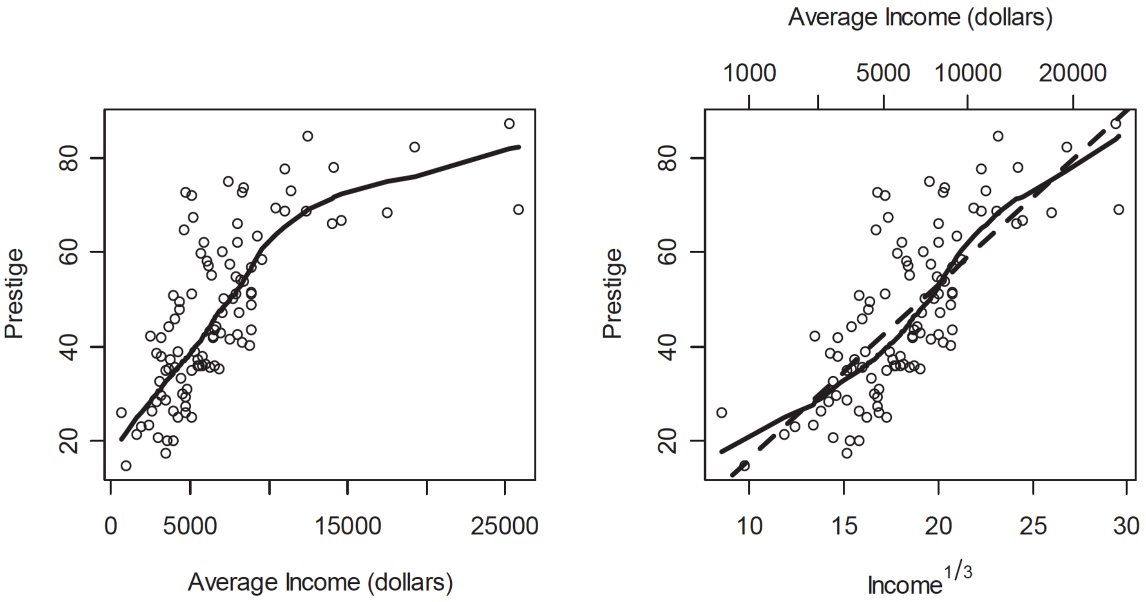


Figure 8: Figure from Fox

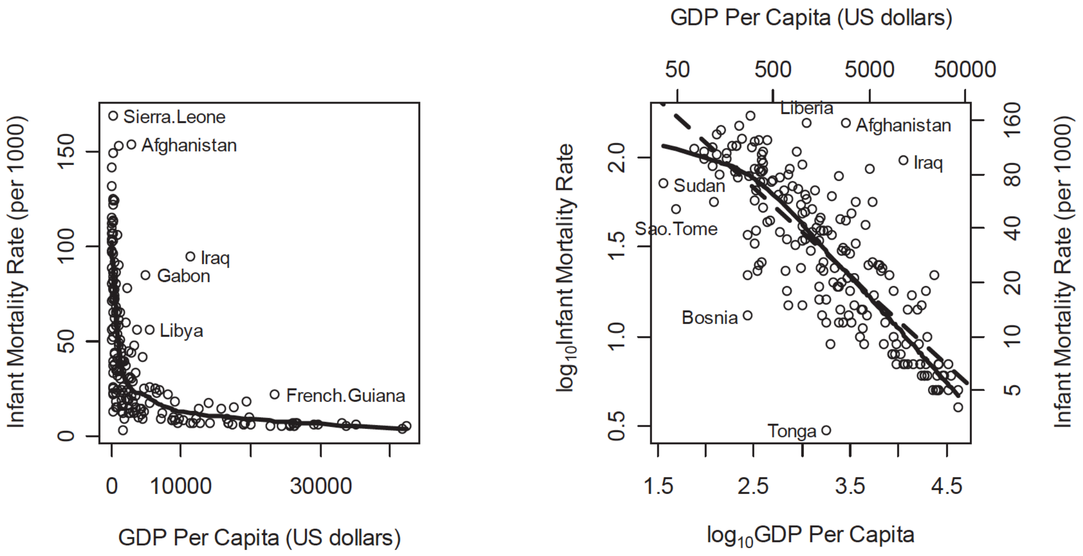


Figure 9: Figure from Fox

the linear fit as log-spread $\approx a + b$ log-level. Then the corresponding spread-stabilizing transformation uses the power $p = 1 - b$.

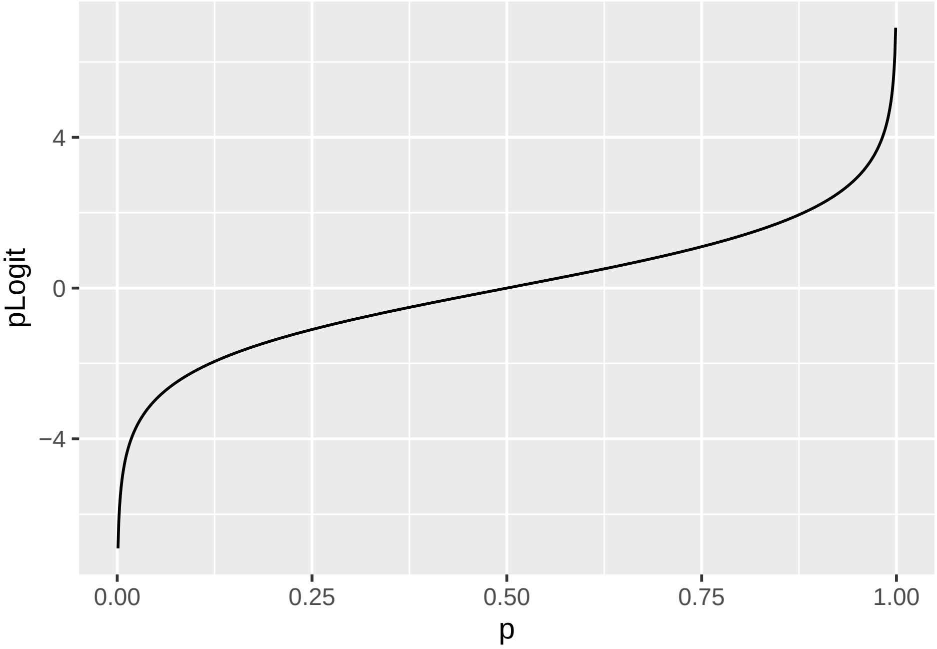
The problems of unequal spread and skewness commonly occur together, because they often have a common origin such as frequency counts: the impossibility of obtaining a negative count tends to produce positive skewness, together with a tendency for larger levels to be associated with larger spreads.

Transforming proportions Power transformations are often not helpful for proportions, since these quantities are bounded below by 0 and above by 1. If the data values do not approach these two boundaries, then proportions can be handled much like other sorts of data. Percents and many sorts of rates are simply rescaled proportions. Several transformations are commonly employed for proportions; the most important is the logit transformation:

$$\text{logit } P = \log_e \frac{P}{1-P}$$

The logit transformation is the log of the odds ratio, $P/(1-P)$

The logit transformation removes the upper and lower boundaries of the scale, spreading out the tails of the distribution and making the resulting quantities symmetric about 0. The transformation is nearly linear in its center, between about $P = 0.2$ and $P = 0.8$.



Interpreting Regression Coefficients after Transformations

Transforming the predictor

The regression model becomes

$$Y = \beta_0 + \beta_1 f(X_1) + \epsilon$$

where f is an invertible, non-linear function. The interpretations still apply but now to $f(X)$, not X. Thus, $\beta_0 = E[Y|f(X) = 0)]$ which is not the same as $E[Y|X = 0]$. β_1 is the slope in units of Y per units of $f(X)$.

That is, it is the difference in the expected response for a difference in $f(X)$ of 1, **not** for a difference in X of 1.

For example, suppose that $f = \log$, then the model is

$$Y = \beta_0 + \beta_1 \log X + \epsilon$$

In this case, $\log X = 0$ means that $X = 1$, so $\beta_0 = E[Y|X = 1]$.

A k unit change in $\log X$ means multiplying X by e^k so β_1 is the expected difference in Y for an e-fold change in X.

Transforming the response

The regression model becomes

$$g(Y) = \beta_0 + \beta_1 X + \epsilon$$

for an invertible, nonlinear function g. The interpretations now apply to $g(Y)$, not Y, because the model for Y is now

$$Y = g^{-1}(\beta_0 + \beta_1 X + \epsilon)$$

Thus, $\beta_0 = E[g(Y)|X = 0]$. Since g is not a linear function, neither is g^{-1} and therefore, $E[Y|X = 0] \neq g^{-1}(\beta_0)$ and more generally, $E[Y|X = x] \neq g^{-1}(\beta_0 + \beta_1 X)$.

β_1 is the difference in the mean of $g(Y)$ predicted by a one-unit change in X. There is generally no simple interpretation of β_1 for the original Y.

In the special case that $g = \log$, then the model is

$$\log Y = \beta_0 + \beta_1 X + \epsilon$$

then

$$Y = e^{\beta_0 + \beta_1 X + \epsilon} = e^{\beta_0} e^{\beta_1 X} e^{\epsilon}$$

e^{β_0} is the median value of Y when $X = 0$.

A one-unit increment in X predicts that Y should be larger by a factor of e^{β_1}. Thus, additive, equal-sized changes to X lead to multiplicative changes in Y.

Although $\epsilon \sim N(0, \sigma^2)$, e^{ϵ} is not Gaussian, rather it follows the *log-normal* distribution (i.e., $e^{\epsilon} \sim LN(0, \sigma^2)$)

Recommendation

If your outcome variable is a count, use Poisson regression. If your outcome variable is a proportion (i.e., binary), then use logistic regression. In other words, instead of transforming your variables, use a model appropriate for the distribution of the outcome variable, if possible, because interpretion will be much easier.