POINTS OF SIGNIFICANCE

Analyzing outliers: influential or nuisance?

Some outliers influence the regression fit more than others.

In our recent columns, we discussed how linear regression can be used to predict the value of a response variable on the basis of one or more predictor variables[1,2]. We saw that even when a fit can be readily obtained, interpreting the results is not as straightforward. For example, when predictors are correlated, regression coefficients cannot be reliably estimated—and may actually have the wrong sign—even though the model remains predictive[2]. This month we turn to methods that diagnose the regression, beginning with the effect that outliers have on the stability of predicted values. Other diagnostics, such as for stability of the regression coefficient estimates and for statistical inference, will be the subject of a future column.

Recall that simple linear regression is a model for the conditional mean $E(Y|X) = \beta_0 + \beta_1 X$ of the response, Y, given a single predictor, X. Because of biological or technical variability, we expect deviation between the conditional mean and the observed response. This is called the error, and when it can be assumed to be additive, be independent and have zero mean, least-squares estimation (LSE) is most commonly used to determine the respective estimates b_0 and b_1 of regression parameters β_0 and β_1. LSE minimizes the residual sum of squares, $SSE = \Sigma(y_i - \hat{y}_i)^2$, where $\hat{y}_i = b_0 + b_1 x_i$ are the fitted values. An estimate of the error is given by the residual $r_i = y_i - \hat{y}_i$. In addition, it is often assumed that errors are normally distributed and have constant variance that is independent of the values of the predictors.

One of the most common regression diagnostics involves identifying outliers and evaluating their effect on the estimates of the fit parameters; this helps us understand how much influence individual observations have on the fit. To illustrate, we will use our simple linear regression model[1] that relates height (H, in centimeters) to weight (W, in kilograms): $W = -45 + 2H/3 + \varepsilon$, with ε normally distributed with zero mean and $\mathrm{Var}(\varepsilon) = 1$.

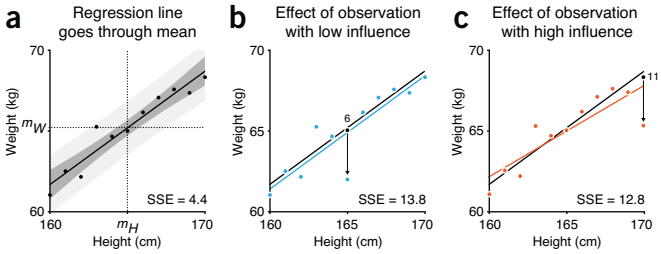

Figure 1 | Observations near the mean have less influence on the regression estimates and fitted values. (**a**) A simple linear regression line always goes through the means of the predictor and the response. Shown are values for a sample with $n = 11$ (black dots) simulated with $W(H) = -45 + 2/3H + \varepsilon$, with the noise distributed normally and with zero mean and variance of 1. The regression (black line) passes through (mean height $m_H = 165$, mean weight $m_W = 65.2$) and has a slope of 0.70. Also shown are the 95% confidence interval (dark gray band) and 95% prediction interval (light gray band). (**b**) The fit (blue line) for a new sample (blue dots) derived from observations shown in **a** by modifying the sixth weight at $H = 165$ from $W(165) = 65$ to $W(165) = 62$. The black line is the fit from **a**. (**c**) Same as **b**, except here we obtained the new sample (orange dots) by changing the 11th weight in **a** from $W(170) = 68.3$ to $W(170) = 65.3$. The sum of squared residuals (SSE) is shown for each fit.

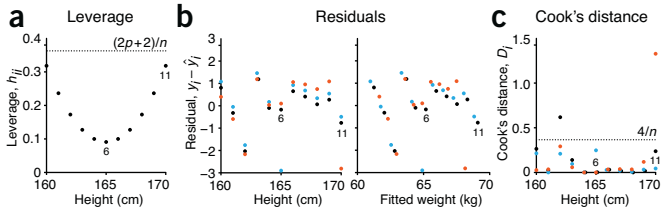

Figure 2 | The leverage, residual and Cook's distance of an observation are used to assess the robustness of the fit. (**a**) The leverage of an observation tells us about its potential to influence the fit and increases as the square of the distance from the predictor value to its mean. Shown are leverage values for the data set in **Figure 1a**. Leverage values larger than $(2p + 2)/n = 0.36$ (dotted line; $p = 1$, $n = 11$) are considered large for p predictors and sample size n. (**b**) The residual is the distance between the observation and its fitted value, $y_i - \hat{y}_i$, shown here for the three fits in **Figure 1** as a function of the predictor value (left) and fitted weight (right). Colors of points correspond to the colors of the fitted lines in **Figure 1**, and there is a horizontal offset of half the width of a data point where points would otherwise occlude each other. (**c**) Cook's distance is a measure of the influence of each data value on the fit and values greater than $4/n = 0.36$ (dotted line; $n = 11$) are considered high influence. Shown are Cook's distances for each fit in **Figure 1**.

A key observation is that the regression line always goes through the predictor and response mean (**Fig. 1a**). The means act as a pivot, and if the predictor value is far from the mean, any unusual values of the corresponding response lead to larger 'swings' in the regression slope. As a consequence, observations farther from the mean have a greater effect on the fit. We show this in **Figure 1b**,**c**, where we simulate an outlier by subtracting three times the noise in the model, $3\mathrm{Var}(\varepsilon)$, from an observation in the sample shown in **Figure 1a**. Subtracting from the sixth observation has very little impact on the fitted value at this position, which drops from 65.2 to 64.9, and essentially no effect on the slope (**Fig. 1b**). Doing the same to the 11th observation decreases both to a greater extent: the fitted value drops from 68.7 to 67.5, and the slope from 0.70 to 0.57 (**Fig. 1c**).

Note that this adjustment also affects the SSE, which is used to estimate the standard errors of the regression coefficients and fitted values, and may have a large effect on the statistical inference even when the influence on the fit is small. For our example, the SSE is larger for the fit obtained by moving the low-leverage observation (**Fig. 1b**) than for the case of the high-leverage one (**Fig. 1c**).

Influence of an observation (x_i, y_i) on the fit can be quantified on the basis of the extent to which a change in the observation affects the corresponding fitted value $\hat{y}_i$. There are two components to influence. The first is due to the distance between x_i and the mean of x, called the leverage, which can be thought of as the effect of a unit change in y_i on the fitted value. The second is due to the distance between y_i and the fitted value at x_i when the line is fitted without (x_i, y_i), captured by a quantity called Cook's distance.

For simple linear regression, the leverage is given by $h_{ii} = 1/n + (x_i - \bar{x})^2/S_{xx}$, where $S_{xx} = \Sigma_i(x_i - \bar{x})^2$ (**Fig. 2a**). The subscript ii

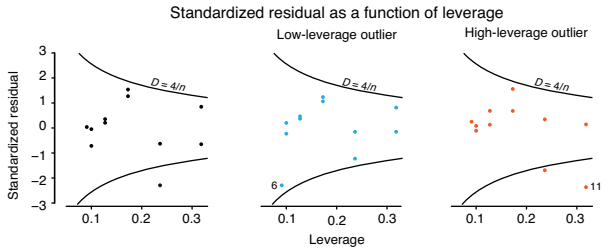

Figure 3 | A plot of residuals as a function of leverage identifies influential observations that are not modeled well by the regression. These quantities are shown here for each of the fits in **Figure 1**. The contour of Cook's distance of $4/n = 0.36$ ($n = 11$) is shown by a black line. The sixth observation that was adjusted (**Fig. 1b**) stands out as a low-leverage outlier (middle panel). In contrast, the 11th observation (**Fig. 1c**) has high leverage, a large residual and a large Cook's distance (right panel).

originates from the fact that the leverage is a diagonal element in the so-called hat matrix. Leverage is minimum when $x_i = \bar{x}$, but not zero—leverage is always between $1/n$ and 1, and an observation affects the fitted value even if it has minimum leverage. Typically an observation is said to have high leverage if $h_{ii} > (2p + 2)/n$. For our example this cutoff is 0.36 for $p = 1$ predictors and a sample size of $n = 11$.

For multiple regression, the computation of h_{ii} is more complicated, but it still measures the distance between the vector of predictors and their mean. It is possible for predictors to individually have typical values but have large h_{ii}. For example, if height and weight are predictors in a sample of adult humans, 55 kg might be a typical weight and 185 cm might be a typical height, but a 55-kg individual who is 185 cm tall could be unusual, and so this particular combination of height and weight can have large leverage.

Recall that fitted values are chosen to minimize the residuals, as the LSE minimizes SSE = Σr_i^2. Thus, because observations with high leverage have greater potential to influence the fit, they can pull the fit toward them and have small residuals, at the cost of increased residuals for low-leverage observations. This can be diagnosed by a plot of the residuals versus the fitted values (**Fig. 2b**). Typically, high-leverage points that also have large error pull the fit away from the other points, creating a trend in the residuals. For our example, the residual of the 11th observation is still large because decreasing it further would increase the magnitude of the other residuals; however, outliers at even higher leverages may have residuals that are smaller than more typical observations. In contrast, outliers in y with small leverage values will appear as large residuals near the center of the plot. In addition to telling us something about the influence of an observation, residuals are useful in identifying lack of fit and assessing the validity of assumptions about the noise, as we will show in a future column.

The leverage of an observation and its residual are different attributes, but both contribute to the observation's influence on the fit. Therefore, it is useful to combine them into a quantity called Cook's distance (**Fig. 2c**), $D_i = (r_i^2/((p + 1) \times \text{MSE})) \times (h_{ii}/(1 - h_{ii})^2)$, where the mean squared error MSE = $\Sigma r_i^2/(n - p - 1)$. Another way to write Cook's distance is $\Sigma_j(\hat{y}_j - \hat{y}_{j(i)})^2/((p + 1) \times \text{MSE})$, where $\hat{y}_{j(i)}$ are the fitted values obtained by excluding observation i. When expressed in this way, Cook's distance can be seen more intuitively as proportional to the distance that the predicted values would move if the observation in question were to be excluded. Thus, Cook's distance is a 'leave one out' measure of how the fitted values (or, equivalently, the slopes) depend on each observation.

The D_i and h_{ii} diagnostics, together with the standardized residual $r_i/\sqrt{\text{MSE}}$, are often considered separately, even though they are related. Large values of any of these indicate that the predicted values and estimated regression coefficients may be sensitive to a few unusual data values. Plots that combine these values can provide information-dense diagnostics, but care is required in their interpretation. For example, the standardized residual can be plotted as a function of leverage (**Fig. 3**). Observations with high leverage and large residuals immediately stand out. However, as mentioned, the fit may be pulled toward outliers with high leverage, resulting in small residuals.

Once outliers have been identified, it remains to be determined how to proceed. If the outliers can be attributed to spurious technical error, handling them may be as simple as removing them from the sample or repeating the experiment. However, they may have arisen purely by chance and be a result of biological variability. In this case, removing them would lead to underestimation of the variability in the data and unduly influence inference. As multiple linear regression is often just a local approximation to a nonlinear process, influential high-leverage points may also indicate that the linear approximation must be restricted to a smaller region of the predictor space.

Except when the outliers can be clearly identified as due to a mistake in the experiment, it is never appropriate to simply remove them from the analysis. In some cases, it is necessary to enlarge the scope of the model to explain the outliers. In others, the effects of the outliers on the fitted model and the resulting scientific conclusions should be discussed. Although it is sometimes appropriate to consider the model that best fits the bulk of the data, and thus not use the outliers for prediction, the outliers that were removed need to be clearly identified, along with the reasons for not using them.

To understand a predictive model, we need to understand not only the predictions but also how they may be perturbed as new data are observed. Outlying data are often the best indicators of the stability of our predictions; if their exclusion disproportionately alters the fit or sways the outcome of inference, a more complete model may be needed.

Corrected after print 14 April 2016.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 999–1000 (2015).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 1103–1104 (2015).

Naomi Altman is a Professor of Statistics at The Pennsylvania State University. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.

Corrigendum: Analyzing outliers: influential or nuisance?

Naomi Altman & Martin Krzywinski
Nat. Methods **13, 281–282 (2016); published online 30 March 2016; corrected after print 14 April 2016**

In the version of this piece initially published, there were two errors. The equation describing mean squared error (MSE) was incorrect in the PDF file. In the legend for Figure 1a, the stated values for mean height and mean weight were switched. The errors have been corrected in the HTML and the PDF versions of the piece.