POINTS OF SIGNIFICANCE

Regression diagnostics

Residual plots can be used to validate assumptions about the regression model.

So far in our discussion of linear regression, we have seen that the estimated regression coefficients and predicted values can be difficult to interpret[1]. When the predictors are correlated[2], the magnitude and even the sign of the estimated regression coefficients can be highly variable, although the predicted values may be stable. When outliers are present[3], both the estimated regression coefficients and the predicted values can be influenced. This month, we discuss diagnostics for the robustness of the estimates and of the statistical inference—that is, the *t*-tests, confidence intervals and prediction intervals that are computed on the basis of assumptions that the errors are additive, normal and independent and have zero mean and constant variance.

Recall that the linear regression model is $Y = \beta_0 + \Sigma\beta_j X_j + \varepsilon$, where Y is the response variable, $X = (X_1, \ldots, X_p)$ are the p predictor variables, $(\beta_0, \ldots, \beta_p)$ are the unknown population regression coefficients and ε is the error, which is normally distributed with zero mean and constant variance σ^2 (often written as $N(0, \sigma^2)$). The response and predicted value for the ith observation are y_i and $\hat{y}_i$, respectively, and the difference between them is the residual $r_i = y_i - \hat{y}_i$. The variance is estimated by the mean squared error $\text{MSE} = \Sigma r_i^2/(n - p - 1)$ for sample size n and p predictors.

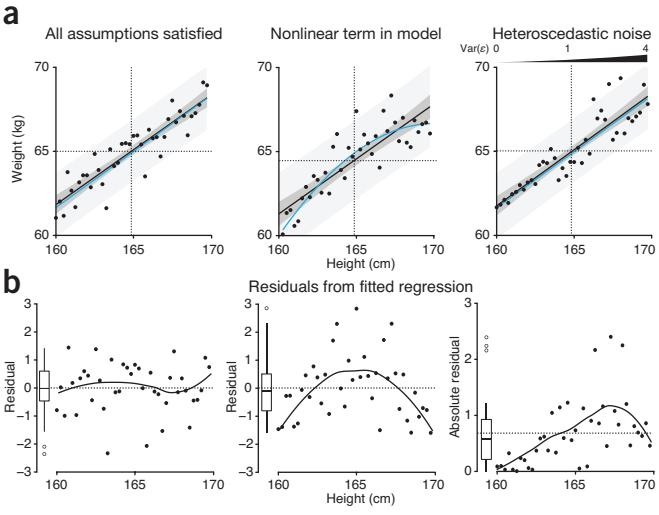

Figure 1 | Residual plots are helpful in assessments of nonlinear trends and heteroscedasticity. (**a**) Fit and residual plot for linear regression of $n = 40$ observations of weight (W) versus height (H) for three scenarios: the linear model $W = -45 + 2H/3 + \varepsilon$, where $\varepsilon \sim N(0, 1)$ (left), the quadratic model $W = -45 + 2H/3 - (H - 165)^2/15 + \varepsilon$, where $\varepsilon \sim N(0, 1)$ (middle), and the linear model with heteroscedastic noise (nonconstant variance) $\varepsilon \sim N(0, ((H - 160)/5)^2)$. Shown are the fit line (black line), model (blue line), sample means (dotted lines), 95% confidence interval (dark gray area) and 95% prediction interval (light gray area). (**b**) Residual plots for the fit including box plots of residuals and smoothed nonparametric fits (solid lines). When assumptions are met, plots should have zero mean, constant spread and no global trends (left). Global trends with zero mean can indicate nonlinear terms (middle). Note that for the heteroscedastic noise scenario, the absolute value of residuals is shown with a mean of 0.68.

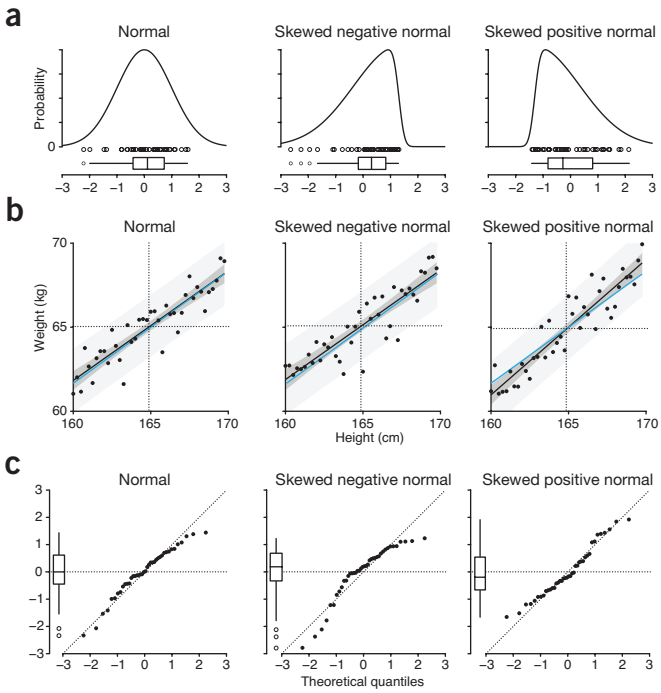

Figure 2 | Q–Q (normal probability) plots compare the differences between two distributions by showing how their quantiles differ. (**a**) Probability plots for $n = 40$ noise samples and their box plots drawn from three noise distributions. The distributions all have means of 0 and variance of 1. (**b**) Regression fits of $n = 40$ observations for the model $W = -45 + 2H/3 + \varepsilon$, where the samples from **a** are used for the noise. Variables and plot elements are defined as in **Figure 1**. (**c**) Q–Q plots and box plots for residuals in fits shown in **b**.

One of the most versatile regression diagnostic methods is to plot the residuals r_i against the predictors (x_i, r_i) and the predicted values $(\hat{y}_i, r_i)$ (**Fig. 1**). When noise assumptions are met, these plots should have zero mean with no local nonrandom trends and constant spread (**Fig. 1b**). Trends indicate that the regression may be nonlinear and that terms such as polynomials (e.g., $\Sigma\gamma_j X_j^2$) may be required in the model for more accurate prediction; residuals will still have an overall zero average (**Fig. 1b**). The absolute values of the residuals can be plotted in the same way to assess the constant variance assumption.

Unless there are substantive reasons to expect a linear relationship between the response and predictor variables, linear regression provides a convenient approximation of the true relationship. However, if the residuals are large and show a systematic trend, there may be a lack of fit between the model and the true relationship (**Fig. 1b**). Whereas linear trends in the residual plots indicate influential data points that have pulled the fit away from the bulk of the data[3], curvature indicates nonlinear trends that have not been captured. Adding powers of the predictors as additional predictors in the model allows us to fit a polynomial, which can capture this curvature. If the polynomial fit exhibits a significantly lower MSE, we might conclude that there are terms present that were not captured in the original model. For example, including an H^2 term in the fit in **Figure 1a** decreases the MSE substantively from 1.32 to 0.86.

A formal test of lack of fit can be done when there are replicates at some combinations of the predictor values. The variability of the

replicates can be used to estimate the error variance—an MSE much larger than the within-replicate variability is evidence that the residuals have an additional component due to lack of fit.

One can assess the assumption of constant noise variance (homoscedasticity) by plotting absolute values of residuals together with a smooth, nonparametric regression line. If the noise is heteroscedastic (nonconstant variance), the plot will have a nonzero mean and the regression line will not be horizontal (**Fig. 1b**).

Although the estimated regression slopes and predicted values are robust to heteroscedasticity, their sampling distributions are not[4]. Most important, the s.d. of the sampling distributions depend on the variances weighted by values of the predictors, so that use of the test statistics and confidence intervals computed under the assumption of constant variance is no longer valid. Prediction intervals will be particularly inaccurate, as they require appropriate coverage of the error distribution as well as of the predicted value. For example, if the data are normally distributed but the variance is not constant (**Fig. 1a**), the MSE will be an estimate of the average variance and the prediction intervals will be too large for values with small variance and too small for values with large variance.

Another factor that can influence the variance estimate is dependence among the errors, which can occur for a number of reasons—for example, multiple observations on the same individual, time or spatial correlation, or a latent random factor that causes familial or other cluster correlations. Error correlation biases the MSE as an estimator of the variance, and that bias can be high and in either direction depending on the type of correlation. As with heteroscedasticity, the estimated regression slopes and predicted values are robust, but their sampling distributions do not have the values computed under the independence assumption[4]. Even if the error variance is constant, in the presence of correlation use of the test statistics, confidence intervals and prediction intervals computed under the independence assumption can lead to very misleading results.

One can detect the correlation of residuals over time by plotting them versus the time at which the observations were made. Ripples and other nonrandom behavior in the plot indicate time correlation. Spatial correlation can similarly be detected in a plot of the residuals versus the spatial coordinates at which the observations were made. Other types of correlation may be more difficult to detect, particularly if they are due to unknown latent variables, such as cluster effects.

Statistical inference for linear regression relies heavily on the variance estimate, MSE, and is therefore influenced by any factor that affects that estimate. Outliers, for example, can increase the MSE in two different ways. Outliers with a large residual, such as low-leverage points, can directly increase the MSE because the MSE is proportional to the sum of squared residuals. Outliers with high leverage and a high Cook's distance[3] may have a small residual but increase MSE indirectly by increasing the residuals of other data points by pulling the linear away from the majority of responses.

Statistical inference is typically done under the assumption that the errors are normally distributed with constant variance. A version of the central limit theorem[5] tells us that for large samples, tests and confidence intervals for the estimated regression coefficients and fitted values continue to be accurate with non-normal data as long as the errors are independent and identically distributed with constant variance. Here, the definition of 'large' depends on the nature of the non-normality. For example, errors that are closer to uniform on a fixed interval are 'close' to normal, but distributions that produce many outliers require large sample sizes. The prediction intervals rely critically on the normality assumption to determine the width and symmetry of the interval. Non-normality of the error distribution may completely disrupt the coverage of prediction intervals.

After it has been established from the residual plots that the residuals have no nonlinear trends and constant variance, informal evaluation of normality is often done using a histogram of the residuals or a Q–Q plot, also known as a normal probability plot (**Fig. 2**). These plots show the sorted values of the sample versus sorted values that would be expected if they were drawn from a normal distribution. Although formal tests of normality can be done, they are not considered to be effective because they may be sensitive to departures from normality that have little effect on the statistical inference.

Correlation among the predictors, also known as multicollinearity, does not affect the stability of the predicted values, but it can greatly affect the stability of the estimated regression coefficients, as we saw in the context of predicting weight from height and maximum jump height[2]. A commonly used measure of multicollinearity is $\mathrm{VIF}(X_i) = 1/(1 - R_i^2)$, where VIF is the variance inflation factor and R_i^2 is the percent variance of X_i explained by the regression of X_i on the other predictors. It is calculated for each predictor X_i via a regression in which values of the predictor are fitted against the other predictors. If $\mathrm{VIF}(X_i)$ is large, then there may be high variation in the regression coefficient estimate between different samples—for example, when VIF > 10, the regression coefficients should not be interpreted. The reciprocal of VIF, the tolerance, is sometimes used equivalently.

Multiple regression is one of the most powerful tools in the basic statistical toolkit. Although simple to apply, it is prone to over- and misinterpretation without attention to diagnostics.

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).
3. Altman, N. & Krzywinski, M. *Nat. Methods* **13**, 281–282 (2016).
4. Eicker, F. in *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1 (eds. Le Cam, L.M., Neyman, J. & Scott, E.M.) 59-82 (Univ. of California Press, 1967)
5. Lumley, T. *et al. Annu. Rev. Public Health* **23**, 151–169 (2002).

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.