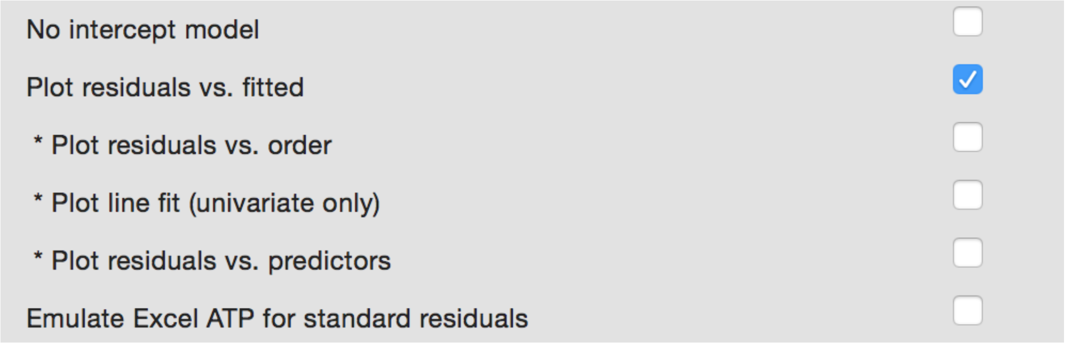
Multiple Linear Regression

The **Multiple Linear Regression** command performs simple multiple regression using least squares. Linear regression attempts to model the linear relationship between variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable (**Response**), and the others are considered to be dependent variables (**Predictors**).

How To

- Run**: Statistics->Regression->Multiple Linear Regression…**
- Select **Dependent** (**Response**) variable and **Independent** variables (**Predictors**)**.**
- To force the regression line to pass through the origin use the **Constant (Intercept) is Zero** option from the **Advanced Options**.
- Optionally, you can add following charts to the report:
  - Residuals versus predicted values plot (use the **Plot Residuals vs. Fitted** option);
  - Residuals versus order of observation plot (use the **Plot Residuals vs. Order** option);
  - Independent variables versus the residuals (use the **Plot Residuals vs. Predictors** option).
- For the univariate model, the chart for the predicted values versus the observed values (**Line Fit Plot**) can be added to the report.



No intercept model ☐
Plot residuals vs. fitted ☑
* Plot residuals vs. order ☐
* Plot line fit (univariate only) ☐
* Plot residuals vs. predictors ☐
Emulate Excel ATP for standard residuals ☐

- Use **the Emulate Excel ATP for standard residuals** option to get the same standard residuals as produced by Excel Analysis Toolpak.

Results

Regression statistics, analysis of variance table, coefficients table and residuals report are produced.

Regression Statistics

R^2 (COEFFICIENT OF DETERMINATION, R-SQUARED) is the square of the sample correlation coefficient between the **PREDICTORS** (independent variables) and **RESPONSE** (dependent variable). In general, R^2 is a percentage of response variable variation that is explained by its relationship with one or more predictor variables.
In simple words, the R^2 indicates the accuracy of the prediction. The larger the R^2 is, the more the total variation of **RESPONSE** is explained by predictors or factors in the model. The definition of the R^2 is

$$R^2 \equiv 1 - \frac{SS_{error}}{SS_{total}}.$$

ADJUSTED R2 (ADJUSTED R-SQUARED) is a modification of R^2 that adjusts for the number of explanatory terms in a model. While R^2 increases when extra explanatory variables are added to the model, the adjusted R^2 increases only if the added term is a relevant one. It could be useful for comparing the models with different numbers of predictors. Adjusted R^2 is computed using the formula:

$Adjusted\ R^2 = 1 - \frac{SS_{error}}{SS_{total}}\frac{df_{total}}{df_{error}} = 1 - (1 - R^2)\frac{N-1}{N-k-1}$, where *k* is the number of predictors excluding the intercept. Negative values (truncated to 0) suggest explanatory variables insignificance, often the results may be improved by increasing the sample size or avoiding correlated predictors.
MSE – the mean square of the error, calculated by dividing the sum of squares for the error term (*residual*) by the degrees of freedom ($df_{error} = n - p, p$ is the number of terms).
RMSE (*root-mean-square error*) – the estimated standard deviation of the error in the model. Calculated as the square root of the MSE.

PRESS – the squared sum of the PRESS residuals, defined in the **RESIDUALS AND REGRESSION DIAGNOSTICS**.
PRESS RMSE is defined as $\sqrt{PRESS/N}$. Provided for comparison with RMSE.
PREDICTED R-SQUARED is defined as $R^2_{Pred} = 1 - PRESS/SS_{Total}$. Negative values indicate that the PRESS is greater than the total SS and can suggest the PRESS inflated by outliers or model overfitting. Some apps truncate negative values to 0.

TOTAL NUMBER OF OBSERVATIONS N - the number of observations used in the regression analysis.

The **REGRESSION EQUATION** takes the form $\mathrm{Y} = a_1x_1 + a_2x_2 + \cdots + a_kx_k + \mathrm{c} + e$, where *Y* is the dependent variable and the *a*'s are the regression coefficients for the corresponding independent terms x_i (or slopes), ***c*** is the constant or ***intercept***, and ***e*** is the error term reflected in the residuals. Regression equation with no interaction effects is often called *main effects model*.
When there is a single explanatory variable the regression equation takes the form of equation of the straight line: $\mathrm{Y} = \mathrm{a}\,x + \mathrm{c}$. Coefficient a is called a slope and c is called an intercept. For this simple case the slope is equal to the correlation coefficient between Y and x corrected by the ratio of standard deviations.

Analysis of Variance Table

SOURCE OF VARIATION - the source of variation (term in the model). The **TOTAL** variance is partitioned into the variance, which can be explained by the independent variables (**REGRESSION**), and the variance, which is not explained by the independent variables (**ERROR**, sometimes called **RESIDUAL**).
SS (SUM OF SQUARES) - the sum of squares for the term.

DF (DEGREES OF FREEDOM) - the number of observations for the corresponding model term. The **TOTAL** variance has $\mathrm{N} - 1$ degrees of freedom. The **REGRESSION** degrees of freedom correspond to the number of coefficients estimated, including the intercept, minus 1.
MS (MEAN SQUARE) - an estimate of the variation accounted for by this term. $MS = SS/DF$
F - the F-test statistic.
P-VALUE - p-value for a F-test. A value less than α level shows that the model estimated by the regression procedure is significant.

Coefficient Estimates

COEFFICIENTS - the values for the regression equation.
STANDARD ERROR - the standard errors associated with the coefficients.
LCL, UCL are the lower and upper confidence intervals for the coefficients, respectively. Default α level can be changed in the **PREFERENCES**.
T STAT - the t-statistics, used to test whether a given coefficient is significantly different from zero.
P-VALUE - p-values for the alternative hypothesis (coefficient differs from 0). A low p-value ($p < 0.05$) allows the null hypothesis to be rejected and means that the variable significantly improves the fit of the model.

VIF – variance inflation factor, measures the inflation in the variances of the parameter estimates due to collinearities among the predictors. It is used to detect multicollinearity problems. The larger the value is, the stronger the linear relationship between the predictor and remaining predictors.
VIF equal to 1 indicates the absence of a linear relationship with other predictors (there is no multicollinearity). VIF value between 1 and 5 indicates moderate multicollinearity, and values greater than 5 suggest that a high degree of multicollinearity is present. It is a subject of debate whether there is a formal value for determining the presence of multicollinearity: in some situations even values greater than 10 can be safely ignored – when high values caused by complicated models with dummy variables or variables that are powers of other variables. In weaker models even values above 2 or 3 may be a cause for concern: for example, for ecological studies Zuur, et al. (2010) recommended a threshold of VIF=3.
TOL - the tolerance value for the parameter estimates, it is defined as *TOL = 1 / VIF*.

Residuals and Regression Diagnostics

PREDICTED values or fitted values are the values that the model predicts for each case using the regression equation.

RESIDUALS are differences between the observed values and the corresponding predicted values. Residuals represent the variance that is not explained by the model. The better the fit of the model, the smaller the values of residuals. Residuals are computed using the formula:

$$e_i = Observed\ value\ -\ Predicted\ value = y_i - \widehat{y_i}.$$

Both the sum and the mean of the residuals are equal to zero.

STANDARDIZED residuals are the residuals divided by the square root of the variance function. Standardized residual is a z-score (standard score) for the residual. Standardized residuals are also known as *standard residuals, semistudentized residuals or Pearson residuals (ZRESID)*. Standardized and studentized residuals are useful for detection of outliers and influential points in regression. Standardized residuals are computed with the untenable assumption of equal variance for all residuals.

$$es_i = \frac{e_i}{s} = \frac{e_i}{\sqrt{MSE}}$$

MSE is the mean squared-error of the model.

STUDENTIZED residuals are the ***internally*** *studentized residuals (SRESID*). The internally studentized residual is the residual divided by *its* standard deviation. The t-score (Student's t-statistic) is used for residuals normalization. The internally studentized residual r_i is calculated as shown below ($\widetilde{h_i}$ is the leverage of the i^th observation).

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1 - \widetilde{h_i})}}$$

DELETED T – *studentized deleted residuals* or **ex*ternally*** *studentized residuals (SDRESID)*, are often considered to be more effective for detecting outlying observations than the internally studentized residuals or Pearson residuals. A rule of thumb is that observations with an absolute value larger than 3 are outliers (Hubert, 2004). Please note that some software packages report the studentized deleted residuals as simply "studentized residuals".

Externally studentized residual t_i (deleted ***t*** residual) is defined as the *deleted residual* divided by its estimated standard deviation.

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_i(1 - \widetilde{h_i})}} = e_i\sqrt{\frac{n - p - 1}{MSE\left(1 - \widetilde{h_i}\right) - {e_i}^2}}$$

p is the number of terms (the number of regression parameters including the intercept).

The *deleted residual* is defined as $d_i = y_i - \widehat{y_{(i)}}$, where $\widehat{y_{(i)}}$ is the predicted response for the i^th observation based on the model with the i^th observation excluded. The mean square error for the model with the i^th observation excluded (MSE_i) is computed from the following equation:

$$(n - p)MSE = (n - p - 1)MSE_i + {e_i}^2/(1 - \widetilde{h_i}).$$

LEVERAGE $\widetilde{h_i}$ is a measure of how much influence each observation has on the model. Leverage of the i^th observation can be calculated as the i^th diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$. Leverage values range from 0 (an observation has no influence) to 1 (an observation has complete influence over prediction) with average value of $\overline{\widetilde{h_i}} = p/n$.

Stevens (2002) suggests to carefully examine *unusual* observations with a leverage value greater than $3p/n$. Huber (2004) considers observations with values between 0.2 and 0.5 as risky and recommends to avoid values above 0.5.

COOK'S D – *Cook's distance*, is a measure of the joint (overall) influence of an observation being an outlier on the response and predictors. Cook's distance expresses the changes in fitted values when an observation is excluded and combines the information of the leverage and the residual. Values greater than 1 are generally considered large (Cook and Weisberg, 1982) and the corresponding observations can be influential. It is calculated as:

$$D_i = \frac{{e_i}^2}{p\ MSE} \frac{\widetilde{h_\iota}}{(1-\widetilde{h_\iota})^2}.$$

DFIT or DFFITS (abbr. *difference in fits*) – is another measure of the influence. It combines the information of the leverage and the studentized deleted residual (*deleted t*) of an observation. DFIT indicates the change of fitted value in terms of estimated standard errors when the observation is excluded. If the absolute value is greater than $2\sqrt{p/(n-p)} \approx 2\sqrt{p/n}$, the observation is considered as an influential outlier (Belsley, Kuh and Welsch, 1980).

$$DFIT_i = t_i \sqrt{\frac{\widetilde{h_\iota}}{1-\widetilde{h_\iota}}}$$

PRESS (**p**redicted **r**esidual **e**rror **s**um of **s**quares) residual or *prediction error* is simply a *deleted residual* defined above*.* The smaller the d_i value is, the better the predictive power of the model. The squared sum of the deleted residuals (PRESS residuals) is known as the *PRESS statistic:* $PRESS = \sum {d_i}^2$.

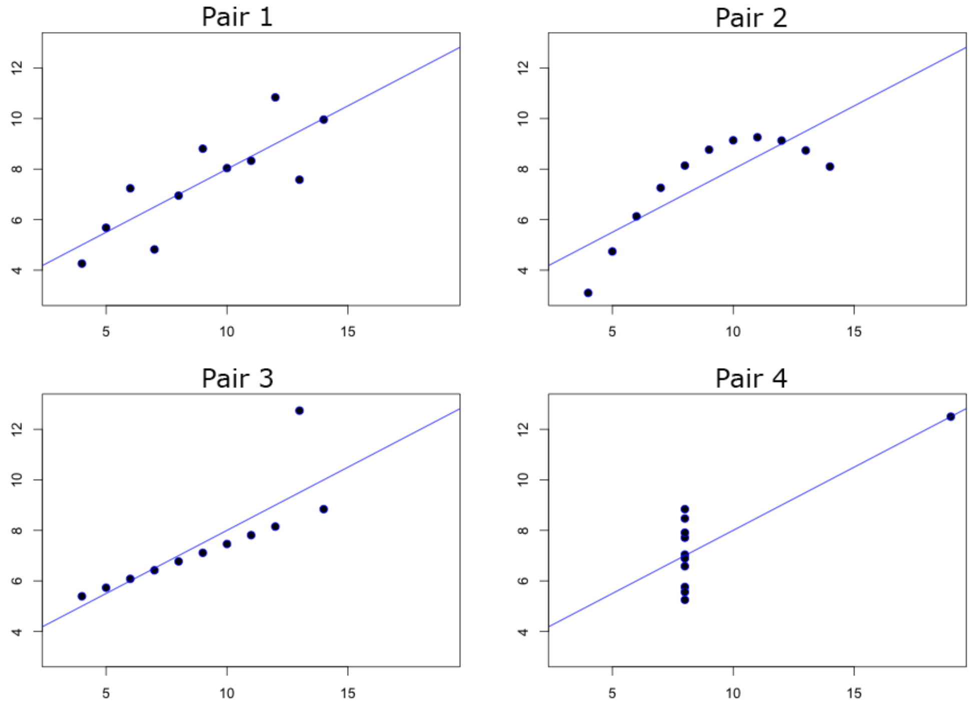
Plots

Unfortunately, even a high R^2 value does not guarantee that the model fits the data well. The easiest way to check for problems that render a model inadequate is to conduct a visual examination.

LINE FIT PLOT

A line fit plot is a scatter plot for the actual data points along with the fitted regression line.

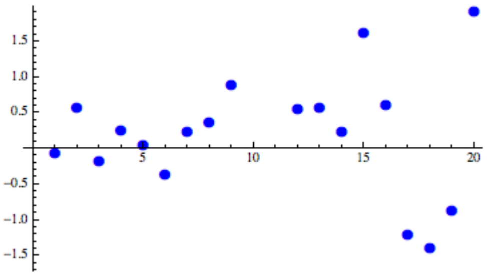
Francis Anscombe demonstrated the importance of graphing data (Anscombe, 1973). The Anscombe's Quartet is four sets of x-y variables [dataset: AnscombesQuartet] that have nearly identical simple statistical properties and identical linear regression equation $y = 3.0 + 0.5x$. *But scatterplots of these variables show that only the first pair of variables has a simple linear relationship and the third pair appears to have a linear relationship except for one large outlier.*



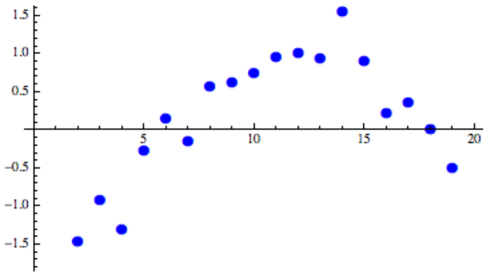
RESIDUAL PLOT

A residual plot is a scatter plot that shows the residuals on the vertical axis and the independent variable on the horizontal axis. It shows how well the linear equation explains the data – if the points are randomly placed above and below x-axis then a linear regression model is appropriate.

Linear model looks appropriate



Nonlinear model would better fit the data



References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27,17-21.

Belsley, David. A., Edwin. Kuh, and Roy. E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons.

Cook, R. Dennis and Weisberg, Sanford (1982). Residuals and Influence in Regression. Chapman and Hall, New York.

Huber, P. (2004). Robust Statistics. Hoboken, New Jersey: John Wiley & Sons.

Neter, J., Wasserman, W. and Kutner, M. H. (1996). Applied Linear Statistical Models, Irwin, Chicago.

Pedhazur, E. J. (1997). Multiple regression in behavioral research (3rd ed.). Orlando, FL: Harcourt Brace.

Stevens, J. P. (2002). Applied multivariate statistics for the social sciences (4th ed.). Mahwah, NJ: LEA.