Multiple linear regression. Several continuous explanatory variables, or covariates, can be measured for each observational unit. If we denote the d covariate measurements for unit i as x_{ij}, $j = 1, \ldots, k$ then the **multiple linear regression model** is defined by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, \ldots, n,$$

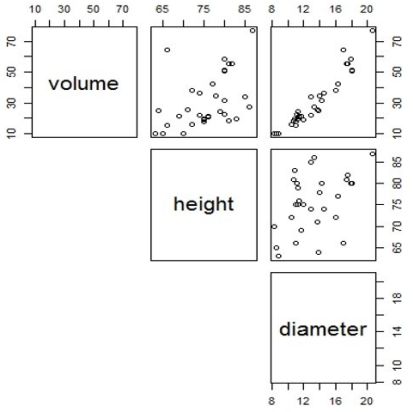
where the residuals are assumed to be independent and normally distributed with mean 0 and variance σ^2. The regression parameters $\beta_1, \ldots, \beta_k$ are interpreted as ordinary regression parameters: a unit change in the variable x_j corresponds to an expected change of β_j in the response y if we assume that all other variables remain unchanged.

Example: Volume of cherry trees. It is difficult to measure the volume of a tree without cutting it down. The tree diameter and height are easy to measure without cutting down the tree, and the primary purpose of this experiment was to predict the tree volume from the diameter and height in order to be able to estimate the value of a group of trees without felling. Thus, the multiple linear regression with two explanatory variables becomes

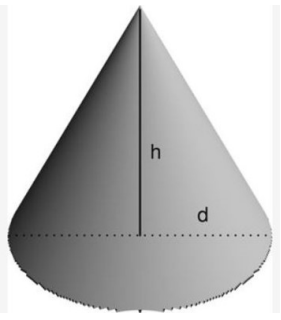
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_k x_{i2} + e_i, \quad i = 1, \ldots, n,$$

where $e_i \sim N(0, \sigma^2)$, and y_i is the volume, x_{i1} is the height, and x_{i2} is the diameter.

We start our initial exploration of the data by producing graphs that show the relationship between the variables. Scatter plot matrix is used to visualize pairwise comparison of variables.



If we assume that the trunk of a tree can be viewed as a cone with diameter d and height h, we can use the result from geometry that gives the volume v of a cone as $v = \frac{\pi}{12} h d^2$



We consider the model

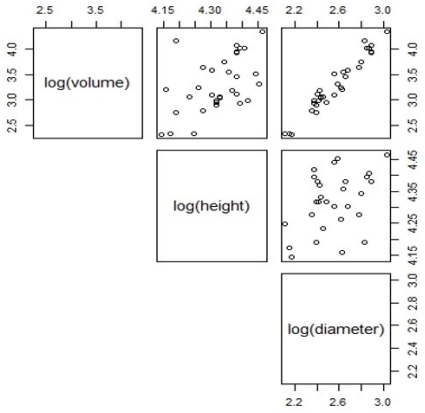
$$v = ch^{\beta_1} d^{\beta_2}$$

with constant c. If the above formula characterizes the volume v better, we may expect $\beta_1 \approx 1$ and $\beta_2 \approx 2$. Thus, in general we get the following model:

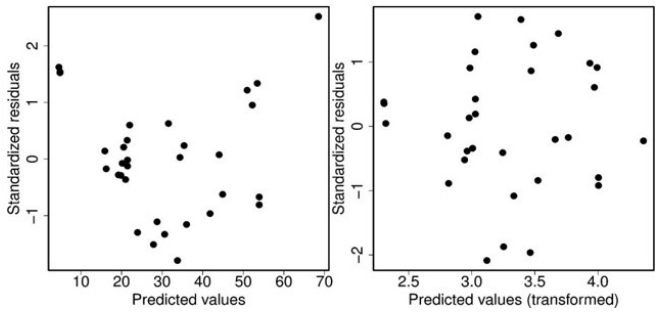
$$\ln y_i = \beta_0 + \beta_1 \ln x_{i1} + \beta_2 \ln x_{i2} + e_i, \quad i = 1, \ldots, n,$$

The model has the form of a multiple linear regression with log-transformed variables, $\ln y$, $\ln x_1$ and $\ln x_2$, corresponding to the volumn y, the height x_1 and the diameter x_2.

Scatter plot matrix is used to visualize pairwise comparison of variables. If we compare these plots to the matrix plot for the original model, the fit has improved.



Residual plots for cherry tree data (left panel) and log-transformed cherry tree data (right panel). The variance in log-transformed data is more homogeneous and there is no apparent structure of the residuals.



Null hypotheses for coefficients. Null hypotheses for the multiple regression model typically corresponds to no effect or no influence of variable x_j on y. The statistical model under the null hypothesis

$$H_0 : \beta_j = 0$$

is a multiple linear regression model without the j-th covariate x_j, but with the other covariates remaining in the model. The test therefore examines if the j-th covariate x_j contributes to the

explanation of variation in y **when the association between** y **and other covariates has been taken into account**.

Example: Volume of cherry trees. In the cherry tree example the hypotheses for $\beta_1 = 0$ and for $\beta_2 = 0$ have been rejected. The conclusion would be that there is association between height and volume and between diameter and volume in the log-transformed data.

Parameter	Estimate	SE	T_{obs}	p-value
β_0	-6.6316	0.7998	-8.292	5.06e-05
β_1	1.1171	0.2044	5.464	7.81e-06
β_2	1.9827	0.0750	26.432	< 2e-16

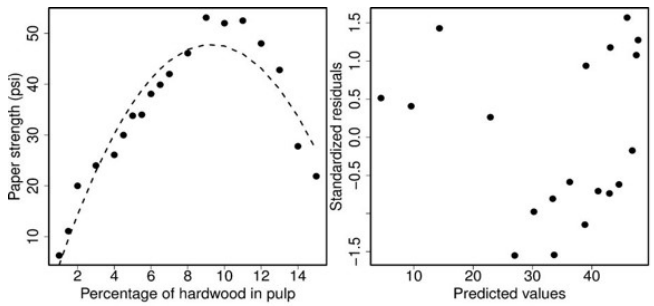
Based on the summary table we conclude that both the height and the diameter are significant and therefore that if we want to model the tree volume, we get the best model when we include information on both diameter and height.

Example: Tensile strength. The data show the tensile strength in pound-force per square inch of Kraft paper (used in brown paper bags) for various amounts of hardwood contents in the paper pulp.

Hardwood	Strength	Hardwood	Strength	Hardwood	Strength
1.0	6.3	5.5	34.0	11.0	52.5
1.5	11.1	6.0	38.1	12.0	48.0
2.0	20.0	6.5	39.9	13.0	42.8
3.0	24.0	7.0	42.0	14.0	27.8
4.0	26.1	8.0	46.1	15.0	21.9
4.5	30.0	9.0	53.1		
5.0	33.8	10.0	52.0		

When we look at the observed strength as a function of hardwood content, it is clear that the strength of the paper starts to decline after the hardwood reaches 11% and that we need a model that is able to capture this change.

Left panel shows paper strength of Kraft paper as a function of hardwood contents in the pulp with the fitted quadratic function superimposed. Right panel is the residual plot for the **quadratic regression model**.



The **quadratic regression model** is given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n$$

This is a special case of the multiple regression model, so we can use the same approach as earlier. In particular, we can test if a quadratic model fits better than a straight line model if we test the hypothesis

$$H_0 : \beta_2 = 0.$$

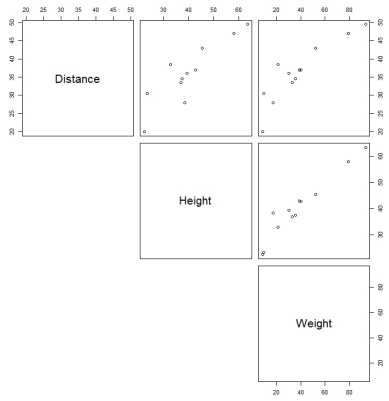
If we reject the null hypothesis we must conclude that the quadratic model fits the data better than the simpler straight line model.

The parameter for the quadratic term, β_2, is highly significant. This tells us that the quadratic regression model is significantly better than a simple linear regression model since we reject the hypothesis $H_0 : \beta_2 = 0$.

Parameter	Estimate	SE	t	p-value
β_0	-6.67419	3.39971	-1.963	0.0673
β_1	11.76401	1.00278	11.731	2.85e-09
β_2	-0.63455	0.06179	-10.270	1.89e-08

Collinearity and multicollinearity. Collinearity is a linear relationship between two explanatory variables and multicollinearity refers to the situation where two or more explanatory variables are highly correlated. For example, two covariates (e.g., height and weight) may measure different aspects of the same thing (e.g., size). Multicollinearity may give rise to spurious results. For example, you may find estimates with the opposite sign compared to what you would expect, **unrealistically high standard errors**, and insignificant effects of covariates that you would expect to be significant. The problem is that **it is hard to distinguish the effect of one of the covariates from the others**. The model fits more or less equally well no matter if the effect is measured through one or the other variable.

Example: Congenital heart defect study. Heart catheterization is sometimes performed on children with congenital heart defects. A Teflon tube (catheter) 3mm in diameter is passed into a major vein or artery at the femoral region and pushed up into the heart to obtain information about the heart's physiology and functional ability. The length of the catheter is typically determined by a physician's educated guess. In a small study involving 12 children, the exact catheter length required was determined by using a fluoroscope to check that the tip of the catheter had reached the pulmonary artery (Weindling, 1977). The patients' heights and weights were recorded. The objective was to see how accurately catheter length could be determined by these two variables.



The scatterplots of all pairs of variables provide a useful visual presentation of their relationships. We will refer to these plots as we proceed through the analysis.

We first consider predicting the length by height alone and by weight alone. The results of simple linear regressions are tabulated below.

Simple Regression with Height

Parameter	Estimate	SE	t	p-value
β_0	12.1240	4.2472	2.855	0.017114
β_1	0.5968	0.1013	5.894	0.000152

Simple Regression with Weight

Parameter	Estimate	SE	t	p-value
β_0	25.63746	2.00421	12.792	1.60×10^{-7}
β_1	0.27727	0.04399	6.303	8.87×10^{-5}

It tests the null hypothesis

$$H_0 : \beta_1 = 0$$

for the simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

These null hypotheses are of no real interest in this problem, but we show the tests for pedagogical purposes. Clearly, null hypothesis would be rejected in these cases. The predictions from both models are similar, and the correlation coefficients are 0.881 and 0.894, respectively.

We next consider the multiple regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

of length on height x_{i1} and weight x_{i2} together, since perhaps better predictions may be obtained by using both variables rather than either one alone. The method of least squares produces the following table.

Parameter	Estimate	SE	t	p-value
β_0	21.0084	8.7512	2.401	0.0399
β_1	0.1964	0.3606	0.545	0.5993
β_2	0.1908	0.1652	1.155	0.2777

Applying t-tests would not lead to rejection of either of the hypotheses

$$H_1 : \beta_1 = 0 \text{ or } H_2 : \beta_2 = 0$$

Yet in the simple linear regressions carried out above, the coefficient β_1 was highly significant. A partial explanation of this is that the coefficient β_1 in the simple regressions and the coefficient β_1 in the multiple regression have different interpretations. In the multiple regression, β_1 is the change in the expected value of the catheter length if height is increased by one unit and weight is held constant. It is the slope of the height that describes the relation of length to height and weight; the large standard error indicates that this slope is not well resolved.

To see why, consider the scatterplot of height versus weight. The method of least squares fits a multiple linear regression to the catheter length values that correspond to the pairs of height and weight values. It should be intuitively clear from the figure that the slope of the fitted linear model is relatively well resolved along the line about which the data points fall but poorly resolved along lines on which either height or weight is constant. Variables that are strongly linearly related, such as height and weight in this example, are said to be **highly collinear**. The plot of height versus weight should serve as a caution concerning making predictions from such a study.