Multiple regression

Introduction

Multiple regression is a logical extension of the principles of simple linear regression to situations in which there are several predictor variables. For instance if we have two predictor variables, X_1 and X_2, then the form of the model is given by:

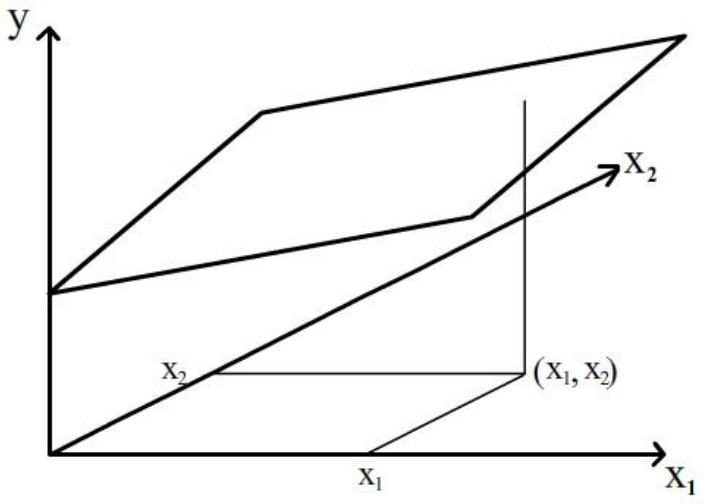
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

which comprises a deterministic component involving the three *regression coefficients* (β_0, β_1 and β_2) and a random component involving the *residual* (error) term, e.

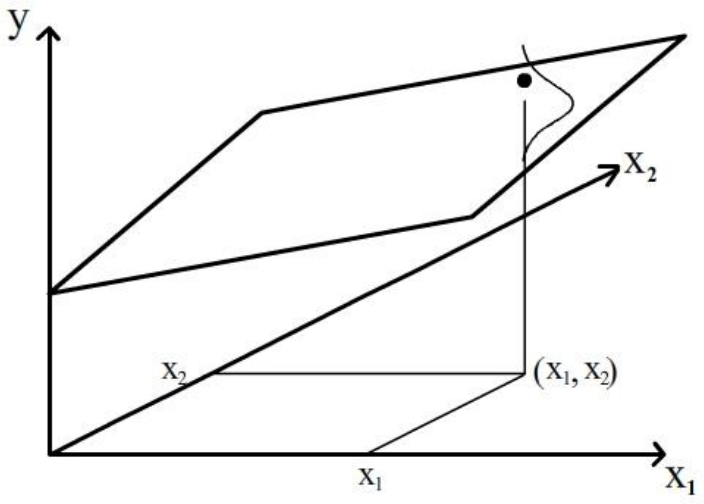
Note that the predictor variables can be either continuous or categorical. In the case of the latter these variables need to be coded as dummy variables (not considered in this tutorial). The response variable must be measured on a continuous scale.

The residual terms represent the difference between the predicted and observed values of individuals. They are assumed to be independently and identically distributed normally with zero mean and variance σ^2, and account for natural variability as well as maybe measurement error.

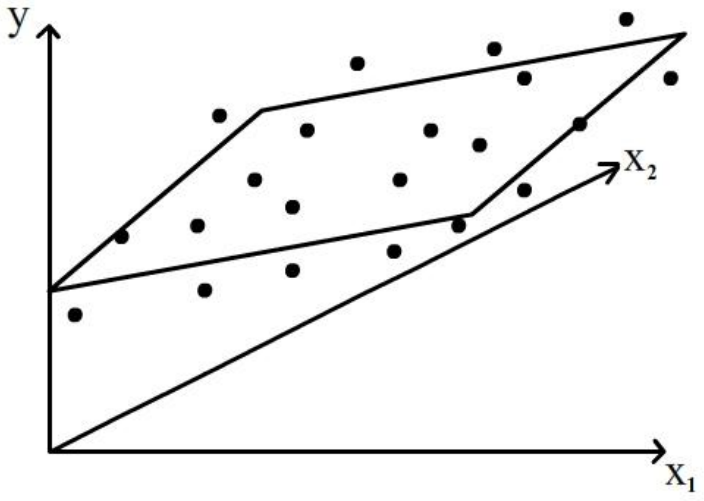
For the two (continuous) predictor example the deterministic component is in the form of a plane which provides the predicted (mean/expected) response for given predictor variable value combinations. Thus if we want the expected value for the specific values x_1 and x_2, then this is obtained from the orthogonal projection from the point (x_1,x_2) in the $X_1 - X_2$ plane to the expected value plane in the 3D space. The resulting Y value is the expected value from this explanatory variable combination.



Observed values for this combination of explanatory variables are drawn from a normal distribution with variance σ^2 centred on the expected value point:



Our data should thus appear to be a collection of points that are randomly scattered with constant variability around the plane.



The multiple regression model fitting process takes such data and estimates the regression coefficients (β_0, β_1 and β_2) that yield the plane that has best fit amongst all planes.

Model assumptions

The assumptions build on those of simple linear regression:

- *Ratio of cases to explanatory variables*. Invariably this relates to research design. The minimum requirement is to have at least five times more cases than explanatory variables. If the response variable is skewed then this number may be substantially more.
- *Outliers*. These can have considerable impact upon the regression solution and their inclusion needs to be carefully considered. Checking for extreme values should form part of the initial data screening process and should be performed on both the response and explanatory variables. Univariate outliers can simply be identified by considering the distributions of individual variables say by using boxplots. Multivariate outliers can be detected from residual scatterplots.
- *Multicollinearity and singularity*. Multicollinearity exists when there are high correlations among the explanatory variables. Singularity exists when there is perfect correlation between explanatory variables. The presence of either affect the intepretation of the explanatory variables effect on the response variable. Also it can lead to numerical problems in finding the regression solution. The presence of multicollinearity can be detected by examining the correlation matrix (say $r = \pm 0.9$ and above). If there is a pair of variables that appear to be highly multicollinear then only one should be used in the regression. Note; some context dependent thought has to be given as to which one to retain!
- *Normality, linearity, homoscedasticity and independence of residuals*. The first three of these assumptions are checked using residual diagnostic plots after having fit a multiple regression model. The independence of residuals is usually assumed to be true if we have indeed collected a random sample from the relvant population.

Example

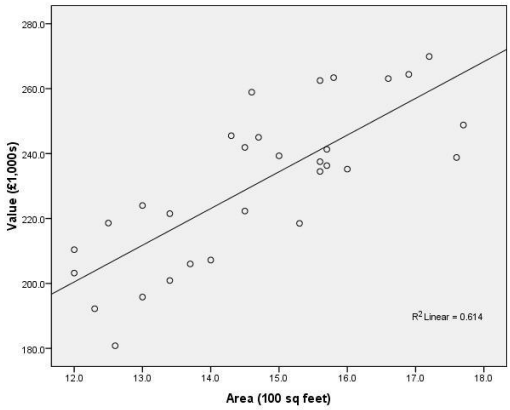
Suppose we are interested in predicting the current market value of houses in a particular city. We have collected data that comprises a random sample of 30 house current values (£1,000s) together with their corresponding living area (100 ft^2) and the distance in miles from the city centre.

Case Summaries

	Value (£1,000s)	Area (100 sq feet)	City centre distance (miles)
1	210.4	12.0	1.2
2	262.5	15.6	1.5
3	258.9	14.6	1.6
4	245.0	14.7	2.5
5	239.3	15.0	2.7
6	263.1	16.6	2.6
7	203.2	12.0	3.2
8	221.5	13.4	3.3
9	207.2	14.0	4.1
10	234.5	15.6	4.2
11	195.8	13.0	4.4
12	222.3	14.5	4.7
13	192.2	12.3	5.1
14	248.8	17.7	5.3
15	218.5	15.3	5.5
16	224.0	13.0	1.3
17	241.9	14.5	1.6
18	245.5	14.3	1.9
19	263.4	15.8	2.4
20	264.4	16.9	2.6
21	269.9	17.2	4.0
22	236.3	15.7	3.2
23	235.2	16.0	4.2
24	218.6	12.5	3.9
25	241.3	15.7	3.8
26	237.5	15.6	4.5
27	180.8	12.6	4.8
28	200.9	13.4	5.0
29	206.0	13.7	5.1
30	238.8	17.6	6.3

Can we build a multiple regression model that can successfully predict house values using the living area and distance variables?

If we consider the relationship between value and area it appears that there is a very significant positive correlation between the two variables (i.e. value increases with area). Fitting a simple linear regression model indicates that 61.4% of the variability in the values is explained by the area.

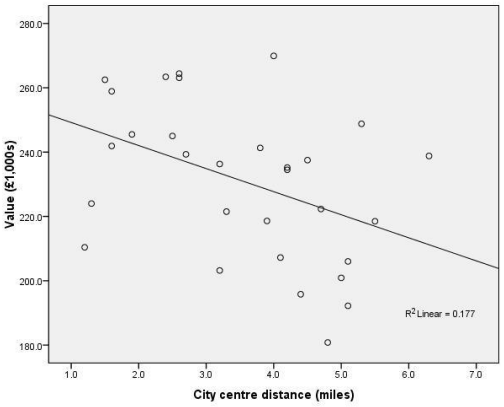


Correlations

		Value (£1,000s)	Area (100 sq feet)
Value (£1,000s)	Pearson Correlation	1	.784**
	Sig. (2-tailed)		.000
	N	30	30
Area (100 sq feet)	Pearson Correlation	.784**	1
	Sig. (2-tailed)	.000	
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

If we consider the relationship between value and distance it appears that there is a significant negative correlation between the two variables (i.e. value decreases with distance). Fitting a simple linear regression model indicates that 17.7% of the variability in the values is explained by the distance from the city centre.



Correlations

		Value (£1,000s)	City centre distance (miles)
Value (£1,000s)	Pearson Correlation	1	-.421*
	Sig. (2-tailed)		.020
	N	30	30
City centre distance (miles)	Pearson Correlation	-.421*	1
	Sig. (2-tailed)	.020	
	N	30	30

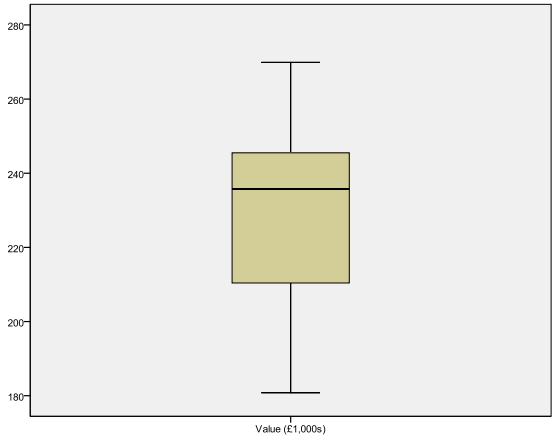
*. Correlation is significant at the 0.05 level (2-tailed).

Thus individually either variable is useful for predicting a house value. We shall now consider the fitting of a multiple regression model that uses both variables for predictions.

First of all we need to address the assumptions that we check before fitting a multiple regression model.

Ratio of cases to explanatory variables.

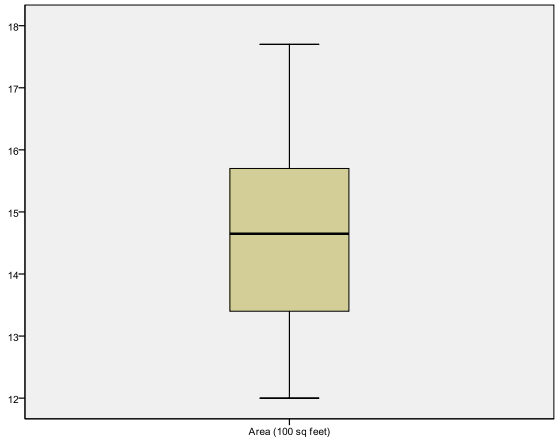
We have 30 cases and 2 explanatory variables. Looking at the boxplot of the response variable value does not overtly worry us that there is a skewness problem.

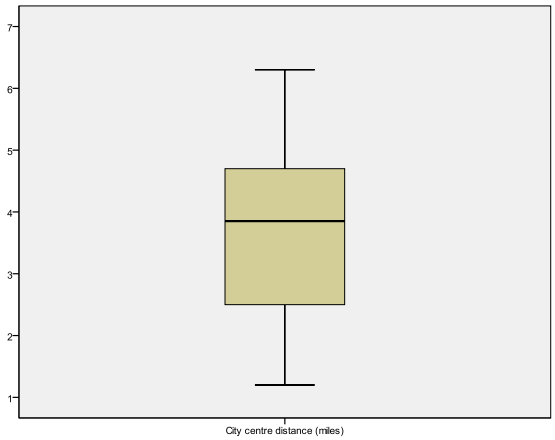


Thus as we have 15 times more cases than explanatory variables we should have an adequate number of cases.

Outliers

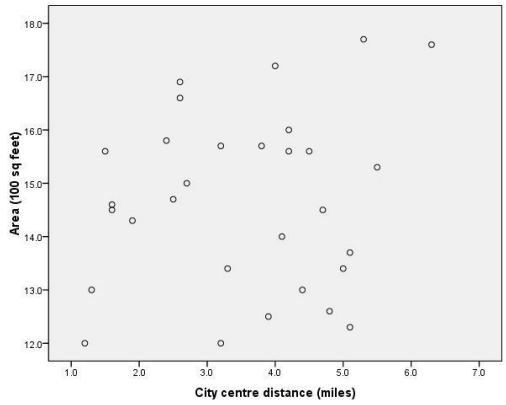
The boxplot above of the response variable does not identify any outliers and neither do the two boxplots below of the explanatory variables:





Multicollinearity and singularity

Examining the correlation between the two explanatory variables reveals that there is not a significant correlation between them. Thus we have no concerns over multicollinearity.



Correlations

		Area (100 sq feet)	City centre distance (miles)
Area (100 sq feet)	Pearson Correlation	1	.154
	Sig. (2-tailed)		.415
	N	30	30
City centre distance (miles)	Pearson Correlation	.154	1
	Sig. (2-tailed)	.415	
	N	30	30

Independence of residuals

Our data has come from a random sample and thus the observations should be independent and hence the residuals should be too.

It appears that our pre model fitting assumption checks are satisfactory, and so we can now consider the multiple regression output.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	80.121	12.051		6.648	.000	55.393	104.848
	City centre distance (miles)	-9.456	.965	-.555	-9.799	.000	-11.436	-7.476
	Area (100 sq feet)	12.548	.818	.869	15.340	.000	10.870	14.226

a. Dependent Variable: Value (£1,000s)

The unstandardized coefficients are the coefficients of the estimated regression model. Thus the expected value of a house is given by:

$$value = 80.121 - 9.456 \times distance + 12.548 \times area.$$

Recalling that value is measured in £1,000s and area is in units of 100 ft^2, we can interpret the coefficients (and associated 95% confidence intervals) as follows.

- For each one mile increase in distance from the city centre, the expected change in house value is -£9,456 (-£11,456, -£7,476). Thus house values drop by £9,456 for each one mile from the city centre.
- For each 100 ft^2 increase in area, the expected house value is expected to increase by £12,548 (£10,870, £14,226).

The significance tests of the two explanatory variable coefficients indicate that both of the explanatory variables are significant ($p<.001$) for predicting house values. If however either had a p-value > .05, then we could infer that the offending variable(s) are not significant for predicting house values.

Note that the intercept here gives the expected value of £80,121 for what would be a house of no area in the exact middle of the city. It is debateable whether this makes any sense and can be dismissed by the fact that these values of the explanatory variables are an extrapolation from what we have observed.

The standardized coefficients are appropriate in multiple regression when we have explanatory variables that are measured on different units (which is the case here). These coefficients are obtained from regression after the explanatory variables are all standardized. The idea is that the coefficients of explanatory variables can be more easily compared with each other as they are then on the same scale. Here we see that the *area* standardised coefficient is larger in absolute value than that of *distance*: thus we can conclude that a change in area has a greater relative effect on house value than does a change in distance from the city centre.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.957[a]	.915	.909	7.2459

a. Predictors: (Constant), Area (100 sq feet), City centre distance (miles)

b. Dependent Variable: Value (£1,000s)

Examining the model summary table:

- The multiple correlation coefficient, R, indicates that we have a very high correlation of .957 between our response variable and the two explanatory variables.
- From the R squared value (*coefficient of determination*) we can see that the model fits the data reasonably well; 91.5% of the variation in the house values can be explained by the fitted model together with the house area and distance from the city centre values.
- The adjusted R square value is attempts to correct for this. Here we can see it has slightly reduced the estimated proportion. If you have a small data set it may be worth reporting the adjusted R squared value.
- The standard error of the estimate is the estimate of the standard deviation of the error term of the model, σ. This gives us an idea of the expected variability of predictions and is used in calculation of confidence intervals and significance tests.

The remaining output is concerned with checking the model assumptions of normality, linearity and homoscedasticity of the residuals. Residuals are the differences between the observed and predicted responses. The residual scatterplots allow you to check:

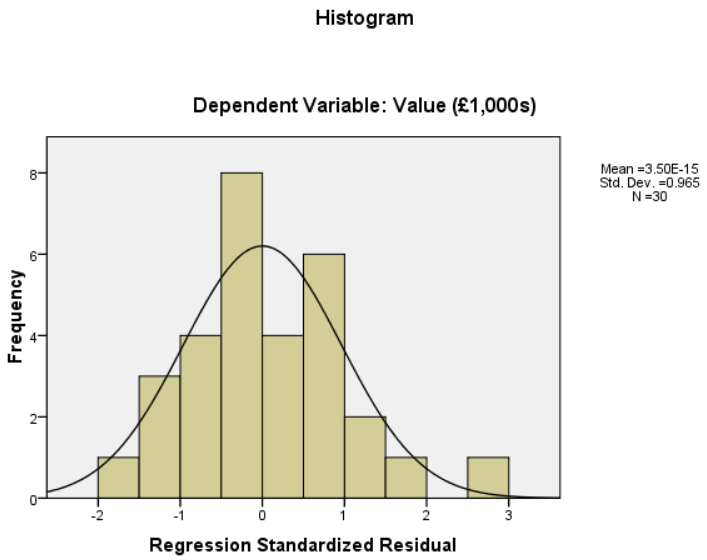
- *Normality*: the residuals should be normally distributed about the predicted responses;
- *Linearity*: the residuals should have a straight line relationship with the predicted responses;
- *Homoscedasticity*: the variance of the residuals about predicted responses should be the same for all predicted responses.

Residuals Statistics[a]

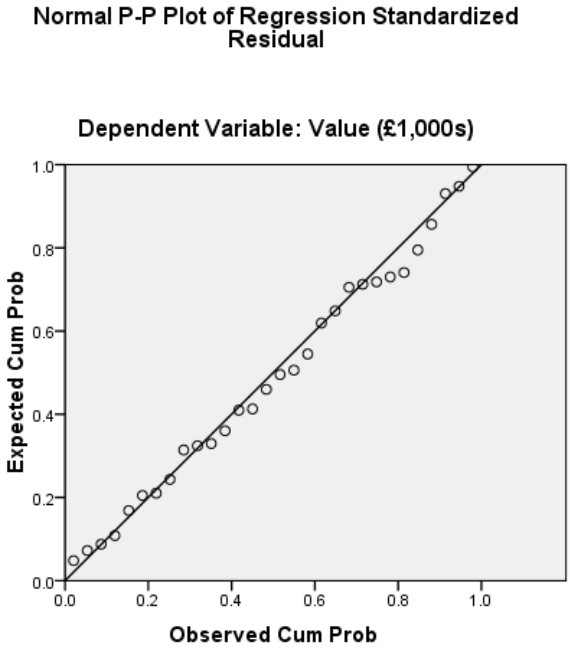
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	186.234	267.596	230.923	22.9872	30
Residual	-12.0357	18.5084	.0000	6.9916	30
Std. Predicted Value	-1.944	1.595	.000	1.000	30
Std. Residual	-1.661	2.554	.000	.965	30

a. Dependent Variable: Value (£1,000s)

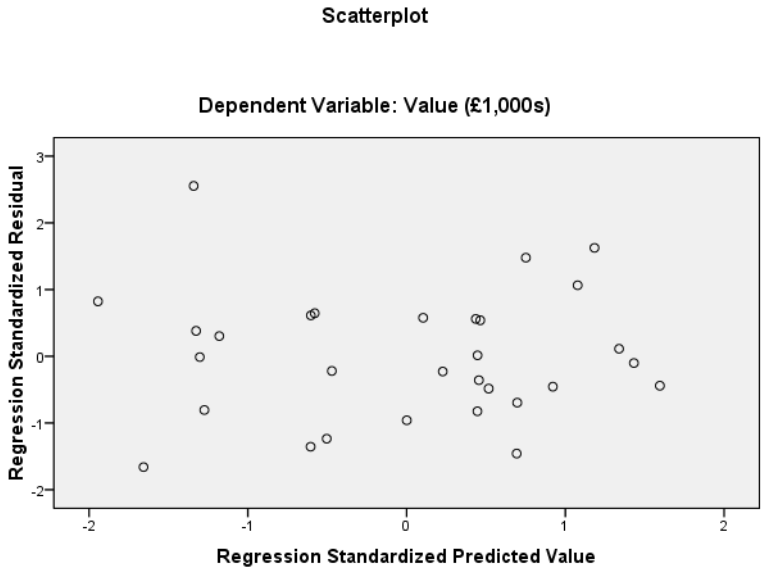
The above table summarises the predicted values and residuals in unstandarised and standardised forms. It is usual practice to consider standardised residuals due to their ease of interpretation. For instance outliers (observations that do not appear to fit the model that well) can be identified as those observations with standardised residual values above 3.3 (or less than -3.3). From the above we can see that we do not appear to have any outliers.



The above plot is a check on normality; the histogram should appear normal; a fitted normal distribution aids us in our consideration. Serious departures would suggest that normality assumption is not met. Here we have a histogram that does look reasonably normal given that we have only 30 data points and thus we have no real cause for concern.



The above plot is a check on normality; the plotted points should follow the straight line. Serious departures would suggest that normality assumption is not met. Here we have no major cause for concern.



The above scatterplot of standardised residuals against predicted values should be a random pattern centred around the line of zero standard residual value. The points should have the same dispersion about this line over the predicted value range. From the above we can see no clear relationship between the residuals and the predicted values which is consistent with the assumption of linearity.

Thus we are happy that the assumptions of the model have been met and thus would be confident about any inference/predictions that we gain from the model.

Predictions

In order to get an expected house *value* for particular *distance* and *area* values we can use the fitted equation. For example, for a house that is 5 miles from the city centre and is 1,400 ft^2:

$$\begin{aligned} value &= 80.121 - 9.456 \times 5 + 12.548 \times 14 \\ &= 208.513 \end{aligned}$$

i.e. £208,513.

Alternatively, we could let a statistics program do the work and calculate confidence or prediction intervals at the same time. For instance, when requesting a predicted value in SPSS we can also obtain the following:

- the predicted values for the various explanatory variable combinations together with the associated standard errors of the predictions;
- 95% CI for the expected response;
- 95% CI for individual predicted responses.

Returning to our example we get the following:

- the expected house value is £208,512 (s.e. = 2,067.6);
- we are 95% certain that interval from £204,269 to £212,754 covers the unknown expected house value;
- we are 95% certain that interval from £193,051 to £223,972 covers the range of predicted individual house value observations.