BMJ

*BMJ* 2013;347:f4373 doi: 10.1136/bmj.f4373 (Published 5 July 2013)



ENDGAMES

STATISTICAL QUESTION

# Multiple regression

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, Tooting, London, UK

Researchers investigated the association between sleep apnoea and hypertension. A total of 2677 adults, aged 2085 years, referred to a sleep clinic in Toronto with suspected sleep apnoea syndrome were recruited. Severity of sleep apnoea was measured using the apnoeahypopnoea index, defined as the total number of apnoeic events plus hypopnoeic events divided by the total number of hours of sleep. Blood pressure measurements were averaged across several readings, taken under standard conditions, with the patients awake and supine, just before getting out of bed in the morning.[1]

Multiple regression was used to examine the association between blood pressure and apnoeahypopnoea index, age, sex, and neck circumference. Separate regression analyses were performed for systolic and diastolic blood pressure (table⇓). Analyses were restricted to 1865 patients not taking antihypertensive drugs.

Which of the following statements, if any, are true?

a) Systolic and diastolic blood pressures were assumed to be linearly related to each of apnoea-hypopnoea index, age, and neck circumference

b) It can be concluded that the effect of the apnoea-hypopnoea index was independently associated with systolic and diastolic blood pressure in patients with apnoea not taking antihypertensive drugs

c) The results of the analyses can be extrapolated outside the observed range of values for apnoeahypopnoea index, age, and neck circumference

d) The results of the analyses can be generalised to all patients referred to the sleep clinic with suspected sleep apnoea

## Answers

Statements *a* and *b* are true, whereas *c* and *d* are false.

The aim of the multiple regression analyses was to predict blood pressure using values of apnoeahypopnoea index, age, sex, and neck circumference. Patients were those referred to a sleep clinic with suspected sleep apnoea. Analysis was restricted to 1865 patients not taking hypertensive drugs. Separate analyses were undertaken for systolic and diastolic blood pressure. The systolic and diastolic blood pressures are referred to as the dependent variables, whereas apnoeahypopnoea index, age, sex, and neck circumference are referred to as the independent, predictor, or explanatory variables.

Multiple regression is an extension of simple linear regression, described in a previous question.[2] Simple linear regression investigates the association between a dependent variable and one explanatory variable. Both the dependent and explanatory variables are continuous, and the association between them is considered to be linear. Multiple regression, sometimes referred to as multivariable analysis, investigates two or more explanatory variables simultaneously, and as in the above example the explanatory variables can be a mixture of continuous and categorical variables. For each regression analysis above, the association between blood pressure and the explanatory variables—apnoeahypopnoea index, age, and neck circumference—was assumed to be a linear one (*a* is true). For the explanatory variable of sex, the analysis provides the difference in the mean value of blood pressure between men and women. If the explanatory variables were all continuous then the analyses would have been called multiple linear regression.

For the multiple regression analysis with systolic and diastolic blood pressure as the dependent variables, the slope or gradient of the linear line with each of the explanatory variables are shown in the column headed β—the so called regression coefficients. The 95% confidence interval (CI) for each coefficient provides an interval estimate for the population parameter of the slope of the linear association. A P value is provided for testing the statistical null hypothesis (in the population from which the patients were sampled the slope of the linear association between the dependent and explanatory variable was zero) against the alternative hypothesis (the slope is not equal to zero). The researchers did not provide the estimated intercepts for the multiple regression analyses—that is, the value of systolic and diastolic blood pressure when all of the explanatory variables are zero. It was therefore not possible to predict blood pressure measurements given values of the explanatory variables, only to describe the nature of any association.

p.sedgwick@sgul.ac.uk

The regression coefficients are sometimes referred to as the estimated partial regression coefficients. They represent the amount by which blood pressure changes on average if the explanatory variable is increased by one unit and all the other explanatory variables are kept constant by being controlled or adjusted for. Therefore, the dependent variable of diastolic blood pressure increased on average by 0.07 mm Hg for each unit increase in the apnoea-hyponoea index, by 0.21 mm Hg for each one year increase of age, and by 0.47 mm Hg for each 1 cm increase in neck circumference. For the regression analysis of diastolic blood pressure, the coefficient for sex (male) was 2.05. Therefore, on average men had a systolic blood pressure that was 2.05 mm Hg higher than women.

For diastolic blood pressure, the regression coefficient for all of the explanatory variables was significantly different from zero, as indicated by a P value less than 0.05 (5%) and a 95% confidence interval that did not include zero. Therefore, the effect of each explanatory variable was said to be independent of all others in the regression analyses. For systolic blood pressure the effects of apnoea-hypopnoea index, age, and neck circumference were independent, whereas that for sex was not. Therefore, the effect of apnoea-hypopnoea index was independently associated with systolic and diastolic blood pressure in patients with apnoea not taking antihypertensive drugs (*b* is true).

The multiple regression lines were calculated using the method of ordinary least squares, often called least squares, as described in a previous question.[2] The analyses made a series of assumptions that are the same as those for simple linear regression. These include, perhaps obviously, that systolic and diastolic blood pressure were linearly associated with each of the continuous explanatory variables— apnoea-hypopnoea index, age, and neck circumference (*a* is true). Secondly, the observations were independent of each other—that is, each patient had only one observation of the dependent and explanatory variables in each of the multiple regression analyses. Thirdly, it was assumed that the residuals in each regression analysis were normally distributed. A residual is the difference between a patient's observed blood pressure value and the predicted value calculated using the regression equation and the patient's apnoeahypopnoea index, age, sex, and neck circumference. It was also assumed that for each linear association between the dependent and explanatory variables, the variation in the dependent variable was consistent for all values of the explanatory variable. If any of the assumptions were in doubt, except for the assumption of independence between observations, then to satisfy the assumptions a transformation of the dependent variable might be considered. A logarithmic transformation, described in a previous question,[3] might be suitable. The analyses should be repeated, including checking the assumptions, using the transformed data.

The multiple regression analyses can be used to predict blood pressure only for the observed range of values of apnoea-hypopnoea index, age, and neck circumference (*c* is false). It is not possible to predict the nature of any association outside the observed ranges of the explanatory variables. The researchers provided this information for all patients referred to the sleep clinic with suspected sleep apnoea syndrome, but not for the patients included in the regression analyses—those patients not taking antihypertensive drugs. Furthermore, the results cannot be generalised to all patients referred to the sleep clinic (*d* is false), particularly those taking antihypertensive drugs. It is not possible to predict the association between blood pressure and the explanatory variables in these patients—it was not investigated and may be very different from that seen for patients not taking antihypertensive drugs. Generalisation and extrapolation of study results has been discussed in a previous question.[4]

Competing interests: None declared.

1 Lavie P, Herer P, Hoffstein V. Obstructive sleep apnoea syndrome as a risk factor for hypertension: population study. *BMJ* 2000;320:479-82.
2 Sedgwick P. Simple linear regression. *BMJ* 2013;346:f2340.
3 Sedgwick P. Log transformation of data. *BMJ* 2012;345:e6727.
4 Sedgwick P. Generalisation and extrapolation of study results. *BMJ* 2013;346:f3022.

Cite this as: *BMJ* 2013;347:f4373

# Table

**Table 1| Multiple linear regression models for blood pressure measurements only in patients not taking antihypertensive drugs (n=1865)**

| Independent variables | Systolic blood pressure | | Diastolic blood pressure | |
|---|---|---|---|---|
| | β (95% CI) | P value | β (95% CI) | P value |
| Apnoea-hypopnoea index (1 apnoeic event) | 0.10 (0.07 to 0.13) | 0.0001 | 0.07 (0.05 to 0.09) | 0.0001 |
| Age (1 year) | 0.39 (0.34 to 0.44) | 0.0001 | 0.21 (0.17 to 0.24) | 0.0001 |
| Sex (male) | −0.70 (−2.50 to 1.11) | 0.45 | 2.05 (0.86 to 3.24) | 0.0007 |
| Neck circumference (1 cm) | 1.01 (0.80 to 1.21) | 0.0001 | 0.47 (0.33 to 0.61) | 0.0001 |