Applied Biostatistics

Multiple Regression

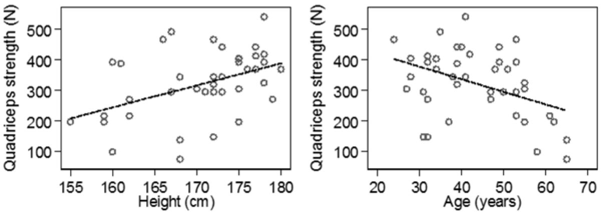
Martin Bland

Professor of Health Statistics
University of York

http://www-users.york.ac.uk/~mb55/msc/

Multiple Linear Regression

More than one predictor:



Strength = –908 + 7.20 × height   Strength = 502 – 4.12 × age

Strength = –466 + 5.40 × height – 3.08 × age

Multiple Linear Regression

More than one predictor:

Strength = –466 + 5.40 × height – 3.08 × age

5.40 is the estimated difference in mean muscle strength between men of the same age who differ in height by one centimetre.

–3.08 is the estimated difference in mean muscle strength between men of the same height who differ in age by one year.

Men who are one year older have muscle strength less by 3.08 newtons.

We say that the 5.40 is the **effect of height adjusted for age**.

Multiple Linear Regression

More than one predictor:

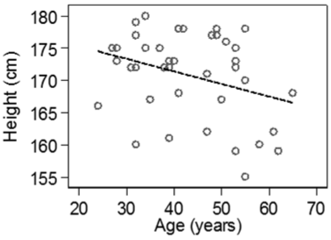
Strength = –908 + 7.20 × height   Strength = 502 – 4.12 × age

Strength = –466 + 5.40 × height – 3.08 × age

Both coefficients are pulled towards zero because age and height are related:

Height = 179 – 0.195 × age,
P = 0.03

Age and height each explains some of the relationship between strength and the other.



Multiple Linear Regression

More than one predictor:

Strength = –466 + 5.40 × height – 3.08 × age
95% CI   0.25 to 10.55   –6.05 to –0.10
                P=0.04   P=0.04

Compare:

Strength = –908 + 7.20 × height
95% CI   2.15 to 12.25
                P=0.006

Strength = 502 – 4.12 × age
95% CI   –7.04 to –1.21
                P=0.007

Each predictor reduces the significance of the other because they are related to one another as well as to strength.

Interactions

Does the age of the subject change the effect of height on strength?

Define an interaction variable.

interaction = height × age

strength = –466 + 5.40 × height – 3.08 × age
P=0.04   P=0.04

strength = 4661 – 24.7 × height – 112.8 × age
+ 0.647 × interaction
P=0.02   P=0.004   P=0.005

If the interaction is significant, both main variables must have a significant effect, so ignore the other P values.

Interactions

Muscle strength data: interaction between height and age.

interaction = height × age

strength = 4661 – 24.7 × height – 112.8 × age
+ 0.647 × interaction
P=0.02 P=0.004 P=0.005

The slope for height depends on age:

slope = –24.7 + 0.647 × age

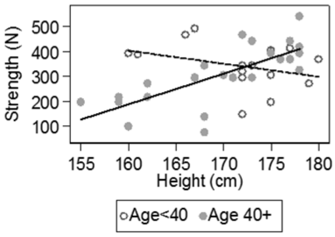
The slope for age depends on height :

slope = –112.8 + 0.647 × height

We cannot interpret the main effects on their own.

Interactions

Muscle strength data: interaction between height and age.

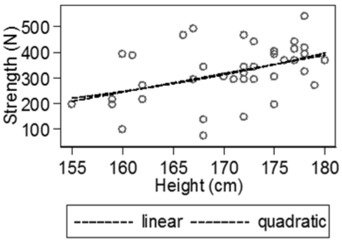


Curvilinear Regression

We can fit a curve:

Strength = 1693 – 23.70 × height + 0.0918 × height^2

Strength = 961 + 7.49 × height + 0.0918 × $(\text{height} – 170)^2$



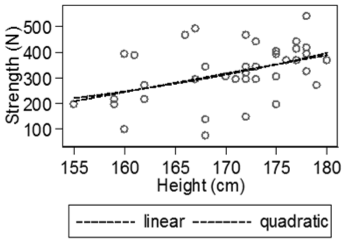
Subtracting a number close to the mean height makes the slope for height easier to interpret.

Curvilinear Regression

We can fit a curve:

Strength = 1693 – 23.70 × height + 0.0918 × height^2

Strength = 961 + 7.49 × height + 0.0918 × (height – 170)^2



In this case the line is almost straight.

height: P = 0.01,

(height – 170)^2: P = 0.8.

Dichotomous predictor variables

Dichotomous predictor: cirrhosis of the liver.

Variable = 1 if subject has cirrhosis, 0 if not.

Strength = –544 + 5.86 × height – 2.75 × age – 34.5 × cirrhosis
P=0.03 P=0.07 P=0.3

Men with cirrhosis have mean strength lower than men without cirrhosis by 34.5 newtons (but not significant, 95% CI for coefficient = –100 to +31).

When we have continuous and categorical predictor variables, regression is also called **analysis of covariance** or **ancova**.

The continuous variables (here height and age) are called **covariates**.

The categorical variables (here cirrhosis) are called **factors**.

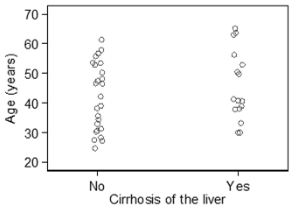
Dichotomous predictor variables

Dichotomous predictor: cirrhosis of the liver.

Variable = 1 if subject has cirrhosis, 0 if not.

Strength = –544 + 5.86 × height – 2.75 × age – 34.5 × cirrhosis
P=0.03 P=0.07 P=0.3

Strength = –466 + 5.40 × height – 3.08 × age
P=0.04 P=0.04



The relationship between cirrhosis and age is sufficient to make age a non-significant predictor of strength.

Multiple correlation coefficient

If we calculate the sum of squares of deviations from the regression line and divide by the sum of squares of the dependent variable about the mean, we get the proportion of variation unaccounted for or not explained by the regression.

One minus this is the proportion of variation explained by the regression, R^2 :

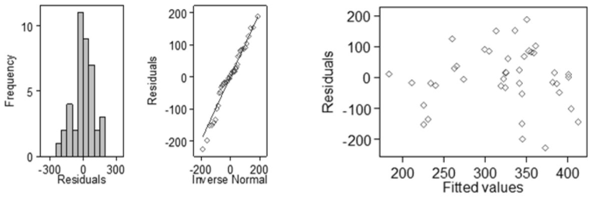
$$R^2 = \frac{\text{SS deviations}}{\text{SS about mean}}$$

R is called the multiple correlation coefficient. Unlike the bivariate correlation, it depends on the choice of dependent variable.

Assumptions of multiple regression

As for simple linear regression, we must assume that the deviations from the regression equation, the differences between the observed values of the outcome variable and the values predicted by the equation, follow a Normal distribution with uniform variance.

Check by plots of the distribution and of deviation against predicted values;



Logistic regression

Logistic regression is used when the outcome variable is dichotomous, a ‘yes or no’.

We want to predict the proportion who have a characteristic, or probability that a subject will have characteristic.

We would like a regression equation:

proportion = intercept + slope × predictor + slope × predictor

Logistic regression

We would like a regression equation:

proportion = intercept + slope × predictor + slope × predictor

Problem: proportions cannot be less than zero or greater than one. How can we stop our equation predicting impossible proportions?

Find a scale for the outcome which is not constrained.

Odds has no upper limit, but must be greater than or equal to zero.

Log odds can take any value.

Use log odds, called the **logit** or **logistic transformation**.

Logistic regression

Example: caesarean section

Several factors may increase the risk of a caesarean, and in this study the factor of interest was obesity, as measured by the body mass index or BMI, defined as weight/height^2 (data of Andreas Papadopoulos).

For caesareans, the mean BMI was 26.7 Kg/m2 and for vaginal deliveries the mean was 24.9 Kg/m2, P < 0.001.

Women who had had a previous vaginal delivery (PVD) were less likely to need a caesarean, odds ratio = 0.18, 95% CI 0.10 to 0.32.

Women whose labour was induced had an increased risk of a caesarean, odds ratio = 1.76, 95% CI 1.19 to 2.62.

Logistic regression

Example: caesarean section

Logistic regression equation:

log odds caesarean =

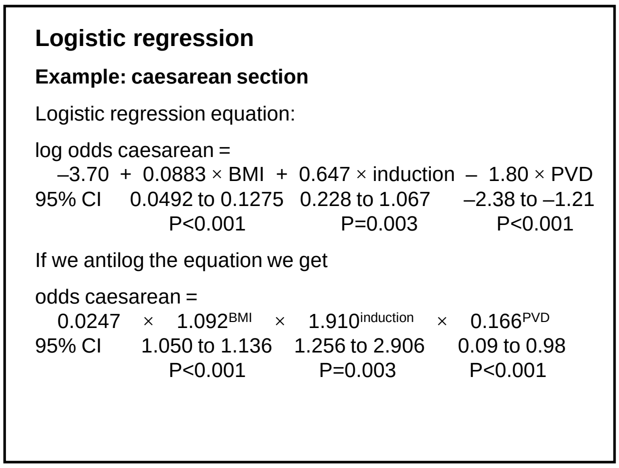
–3.70 +	0.0883 × BMI +	0.647 × induction –	1.80 × PVD
95% CI	0.0492 to 0.1275	0.228 to 1.067	–2.38 to –1.21
	P<0.001	P=0.003	P<0.001

where induction and PVD are 1 if present, 0 if not.

Logistic regression equation predicts log odds.

Coefficients represent the difference between two log odds, a log odds ratio.

The antilog of the coefficients is an odds ratio.



Logistic regression

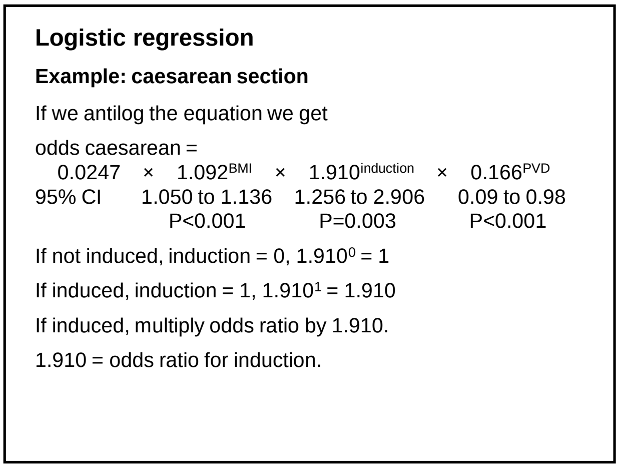
Example: caesarean section

Logistic regression equation:

log odds caesarean =
$-3.70 + 0.0883 \times \text{BMI} + 0.647 \times \text{induction} - 1.80 \times \text{PVD}$
95% CI 0.0492 to 0.1275 0.228 to 1.067 –2.38 to –1.21
$P<0.001$ $P=0.003$ $P<0.001$

If we antilog the equation we get

odds caesarean =
$0.0247 \times 1.092^{\text{BMI}} \times 1.910^{\text{induction}} \times 0.166^{\text{PVD}}$
95% CI 1.050 to 1.136 1.256 to 2.906 0.09 to 0.98
$P<0.001$ $P=0.003$ $P<0.001$



Logistic regression

Example: caesarean section

If we antilog the equation we get

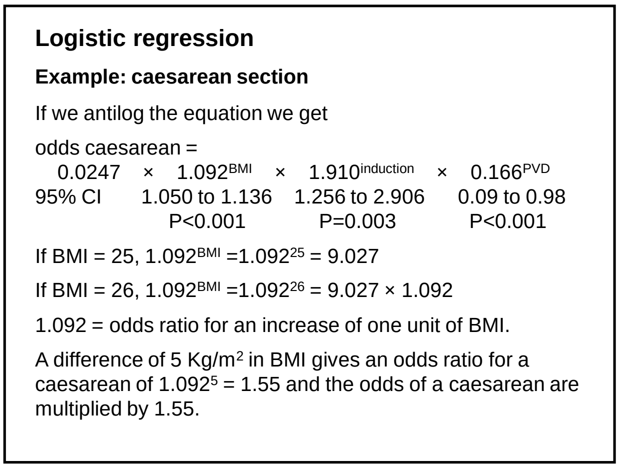
odds caesarean =
$0.0247 \times 1.092^{\text{BMI}} \times 1.910^{\text{induction}} \times 0.166^{\text{PVD}}$
95% CI 1.050 to 1.136 1.256 to 2.906 0.09 to 0.98
$P<0.001$ $P=0.003$ $P<0.001$

If not induced, induction = 0, $1.910^0 = 1$

If induced, induction = 1, $1.910^1 = 1.910$

If induced, multiply odds ratio by 1.910.

1.910 = odds ratio for induction.



Logistic regression

Example: caesarean section

If we antilog the equation we get

odds caesarean =
$0.0247 \times 1.092^{\text{BMI}} \times 1.910^{\text{induction}} \times 0.166^{\text{PVD}}$
95% CI 1.050 to 1.136 1.256 to 2.906 0.09 to 0.98
$P<0.001$ $P=0.003$ $P<0.001$

If BMI = 25, $1.092^{\text{BMI}} = 1.092^{25} = 9.027$

If BMI = 26, $1.092^{\text{BMI}} = 1.092^{26} = 9.027 \times 1.092$

1.092 = odds ratio for an increase of one unit of BMI.

A difference of 5 Kg/m^2 in BMI gives an odds ratio for a caesarean of $1.092^5 = 1.55$ and the odds of a caesarean are multiplied by 1.55.

Factors with more than two levels

We can use factors with more than two levels, i.e. categorical variables with more than two categories as predictors.

Example: follow-up of children of short stature given growth hormone treatment.

Three types of treatment:

1. human growth hormone only (311 children),
2. human growth hormone followed by recombinant growth hormone (1455),
3. recombinant growth hormone only (1467).

Coste J, Letrait M, Carel JC, Tresca JP, Chatelain P, Rochiccioli P, Chaussain JL, Job JC. (1997) Long term results of growth hormone treatment in France in children of short stature: population, register based study. *British Medical Journal* 315, 708-713.

Factors with more than two levels

We can use factors with more than two levels, i.e. categorical variables with more than two categories as predictors.

Example: follow-up of children of short stature given growth hormone treatment.

Three types of treatment:

1. human growth hormone only (311 children),
2. human growth hormone followed by recombinant growth hormone (1455),
3. recombinant growth hormone only (1467).

Hence the treatment is a categorical variable with three categories.

Factors with more than two levels

Three types of treatment:

1. human growth hormone only,
2. human growth hormone followed by recombinant growth hormone,
3. recombinant growth hormone only.

If we code these as 1, 2, and 3, then put this variable as a predictor, the equation is forced to estimate the difference between human growth hormone only and human growth hormone followed by recombinant growth hormone as the same as the difference between human growth hormone followed by recombinant growth hormone and recombinant growth hormone only.

Factors with more than two levels

Define dummy variables, a set of variables which together represent the categorical variable and which can be used in the regression equation.

One way to do this would be:

dummy1 = 1 if human growth hormone only
dummy1 = 0 if any other treatment

dummy2 = 1 if human growth hormone followed by recombinant growth hormone
dummy2 = 0 if any other treatment

We do not need a dummy3, because if dummy1 = 0 and dummy2 = 0, we must have the third treatment, recombinant growth hormone only.

We need one fewer dummy variables than categories.

Factors with more than two levels

Dummy variables:

dummy1 = 1 if human growth hormone only
dummy1 = 0 if any other treatment

dummy2 = 1 if human growth hormone followed by recombinant growth hormone
dummy2 = 0 if any other treatment

Put both dummy variables as predictors into a multiple or logistic regression.

coefficient of dummy1 = difference between human growth hormone only and recombinant growth hormone only.

coefficient of dummy2 = difference between human growth hormone followed by recombinant growth hormone and recombinant growth hormone only.

Factors with more than two levels

Dummy variables:

dummy1 = 1 if human growth hormone only
dummy1 = 0 if any other treatment

dummy2 = 1 if human growth hormone followed by recombinant growth hormone
dummy2 = 0 if any other treatment

The category represented by all the dummy variables being zero is called the reference category, the category to which all the others are compared.

Factors with more than two levels

Coste et al. (1997) chose recombinant hormone only as the reference category and gave the regression coefficients for predicting standard deviation score of final height (i.e. standard deviations from the normal population mean).

For pre-pubertal boys, these were:

human hormone only –0.295 (95% CI –0.456 to –0.134),

human hormone followed by recombinant hormone –0.148 (–0.255 to –0.039),

recombinant hormone only 0.

Coefficient for recombinant hormone only = 0 by definition, because it was the reference category.

Because it is 0 by definition, it has no confidence interval.

Factors with more than two levels

Coste et al. (1997) chose recombinant hormone only as the reference category and gave the regression coefficients for predicting standard deviation score of final height (i.e. standard deviations from the normal population mean).

For pre-pubertal boys, these were:

human hormone only –0.295 (95% CI –0.456 to –0.134),

human hormone followed by recombinant hormone –0.148 (–0.255 to –0.039),

recombinant hormone only 0.

The confidence interval for human hormone only is much wider than for the combination treatment, because there are fewer subjects in this category.

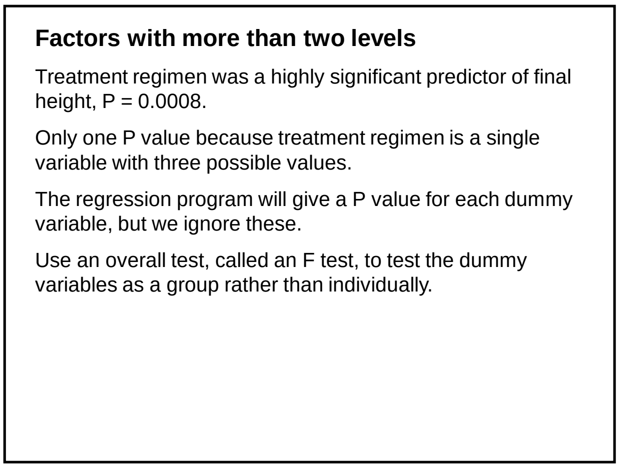
Factors with more than two levels

When we choose the reference category there are two considerations.

We want a category which gives a meaningful comparison. If there is a control group, we usually choose that.

We want a large category so that the confidence intervals would be narrow.

If we had chosen human hormone only as the reference category, though logical, this would mean that both confidence intervals would include comparisons with the small group and so both would be as wide as that between human only and recombinant only.



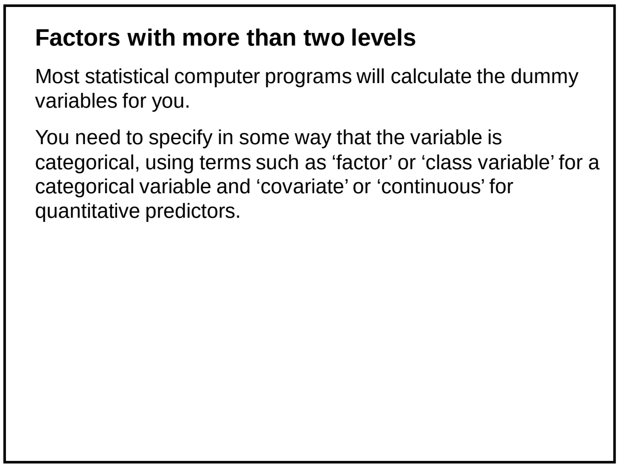
Factors with more than two levels

Treatment regimen was a highly significant predictor of final height, P = 0.0008.

Only one P value because treatment regimen is a single variable with three possible values.

The regression program will give a P value for each dummy variable, but we ignore these.

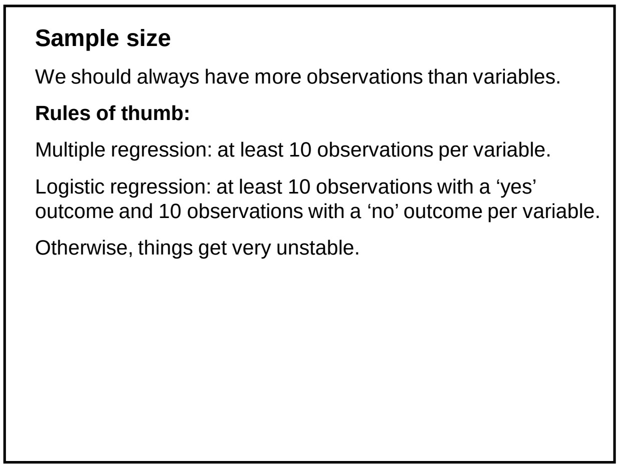
Use an overall test, called an F test, to test the dummy variables as a group rather than individually.



Factors with more than two levels

Most statistical computer programs will calculate the dummy variables for you.

You need to specify in some way that the variable is categorical, using terms such as ‘factor’ or ‘class variable’ for a categorical variable and ‘covariate’ or ‘continuous’ for quantitative predictors.



Sample size

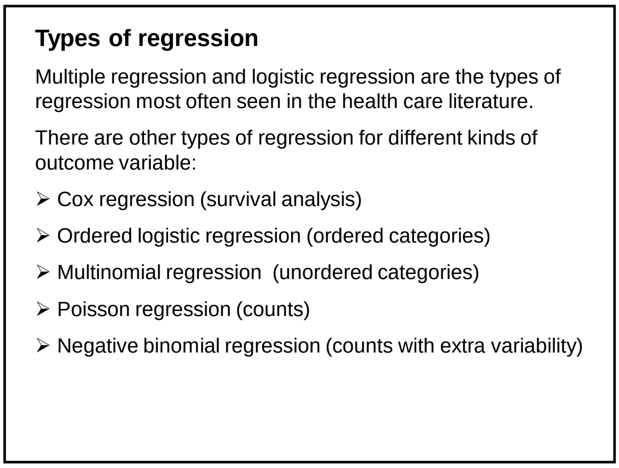
We should always have more observations than variables.

Rules of thumb:

Multiple regression: at least 10 observations per variable.

Logistic regression: at least 10 observations with a ‘yes’ outcome and 10 observations with a ‘no’ outcome per variable.

Otherwise, things get very unstable.



Types of regression

Multiple regression and logistic regression are the types of regression most often seen in the health care literature.

There are other types of regression for different kinds of outcome variable:

- Cox regression (survival analysis)
- Ordered logistic regression (ordered categories)
- Multinomial regression (unordered categories)
- Poisson regression (counts)
- Negative binomial regression (counts with extra variability)