Health Sciences M.Sc. Programme

Applied Biostatistics

Week 10: Multiple regression

More than one predictor

In Week 9 we looked at regression with one predictor variable. Often we would like to use more than one predictor variable. In this lecture we look at how to do that for a continuous outcome variable, and describe a related method for use when the outcome is dichotomous.

Table 1 shows the ages, heights and maximum voluntary contraction of the quadriceps muscle (strength) in a group of male alcoholics. The outcome variable is strength. Figure 1 shows the relationship between strength and height. We can fit a regression line:

$$\text{strength} = -908 + 7.20 \times \text{height}$$

This enables us to predict what the mean strength would be for men of any given height. But strength varies with other things beside height. Figure 2 shows the relationship between strength and age. We can fit a regression line from which we could predict the mean strength for any given age:

$$\text{strength} = 502 - 4.12 \times \text{age}$$

However, strength would still vary with height. To investigate the effect of both age and height, we can use multiple regression to fit a regression equation:

$$\text{strength} = -466 + 5.40 \times \text{height} - 3.08 \times \text{age}$$

The coefficients are calculated by a least squares procedure, exactly the same in principle as for simple regression. In practice, this is always done using a computer program. From this equation, we would estimate the mean strength of men with any given age and height, in the population of which these are a sample.

In this multiple regression equation, 5.40 is the estimated difference in mean muscle strength between men of the same age who differ in height by one centimetre. Similarly, –3.08 is the estimated difference in mean muscle strength between men of the same height who differ in age by one year, i.e. 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**.

Both coefficients are closer to zero than they are in the separate regressions. They are pulled towards zero because, as Figure 3 shows, age and height are related:

$$\text{height} = 179 - 0.195 \times \text{age},$$
$$P = 0.03$$

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

Table 1. Maximum voluntary contraction (strength) of quadriceps muscle, age and height, of 41 male alcoholics (Hickish *et al.*, 1989)

Age (years)	Height (cm)	Strength (newtons)		Age (years)	Height (cm)	Strength (newtons)
24	166	466		42	178	417
27	175	304		47	171	294
28	173	343		47	162	270
28	175	404		48	177	368
31	172	147		49	177	441
31	172	294		49	178	392
32	160	392		50	167	294
32	172	147		51	176	368
32	179	270		53	159	216
32	177	412		53	173	294
34	175	402		53	175	392
34	180	368		53	172	466
35	167	491		55	170	304
37	175	196		55	178	324
38	172	343		55	155	196
39	172	319		58	160	98
39	161	387		61	162	216
39	173	441		62	159	196
40	173	441		65	168	137
41	168	343		65	168	74
41	178	540				

Figure 1. Muscle strength against height

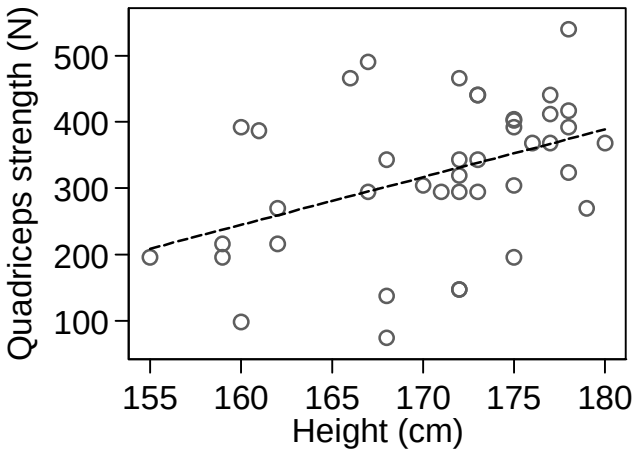


Figure 2. Muscle strength against age

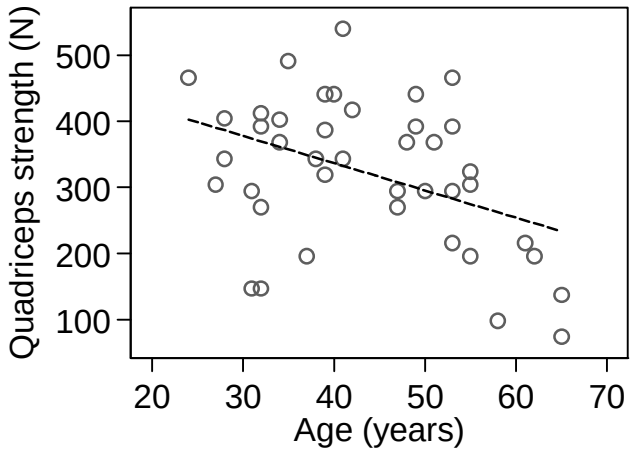
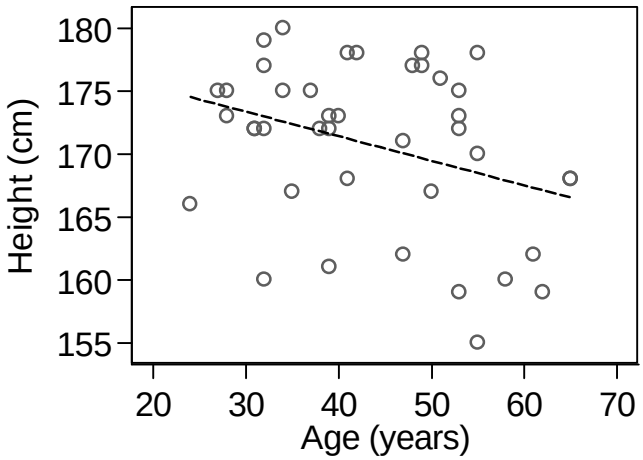


Figure 3. Relationship between height and age in Table 1



Significance tests and estimation in multiple regression

We can test the significance of the regression of strength on height and age together and the significance of each predictor variable separately. These tests and the confidence intervals that go with them require the same assumptions of independent observations and residuals with a Normal distribution and uniform variance as for simple linear regression. For the example, both age and height have P=0.04 and we can conclude that both age and height are independently associated with strength: 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

If we compare this with the separate regressions, we see than the P values have increased:

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. This increases the standard error of the estimates, and variables may have a multiple regression coefficient which is not significant in a multiple regression despite being related to the outcome variable in a simple regression. When the predictor variables are highly correlated the individual coefficients will be poorly estimated and have large standard errors. Correlated predictor variables may obscure the relationship of each with the outcome variable.

We check the assumptions of Normal distribution and uniform variance as for simple linear regression, by plotting a histogram and Normal plot of residuals and scatter plots of the residuals. We usually plot this against the strength predicted by the regression equation.

Interaction in multiple regression

An interaction between two predictor variables arises when the effect of one on the outcome depends on the value of the other. For example, tall men may be stronger than short men when they are young, but the difference may disappear as they age.

An interaction may take two simple forms. As height increases, the effect of age may increase so that the difference in strength between young and old tall men is greater than the difference between young and old short men. Alternatively, as height increases, the effect of age may decrease. If we create an interaction variable = height × age and include it in the model, we can allow either for either of these possibilities:

$$\text{strength} = 4661 - 24.7 \times \text{height} - 112.8 \times \text{age} + 0.650 \times \text{height} \times \text{age}$$
$$\qquad\qquad\qquad\quad P=0.02 \qquad P=0.004 \qquad P=0.005$$

The regression is still significant, as we would expect. However, the coefficients of height and age have changed; they have even changed sign. The coefficient of height depends on age. The regression equation can be written

$$\text{strength} = 4661 + (-24.7 + 0.650 \times \text{age}) \times \text{height} - 112.8 \times \text{age}$$

The coefficient of height depends on age, becoming $-24.7 + 0.650 \times \text{age}$. The difference in strength between short and tall subjects is greater for older subjects than for younger. Or we could write

$$\text{strength} = 4661 - 24.7 \times \text{height} + (-112.8 + 0.650 \times \text{height}) \times \text{age}$$

The coefficient of age depends on height, becoming $-112.8 + 0.650 \times \text{height}$. The difference in strength between young and old subjects being less for taller subjects than for shorter.

Figure 4. Interaction between age and height in their effects on muscle strength

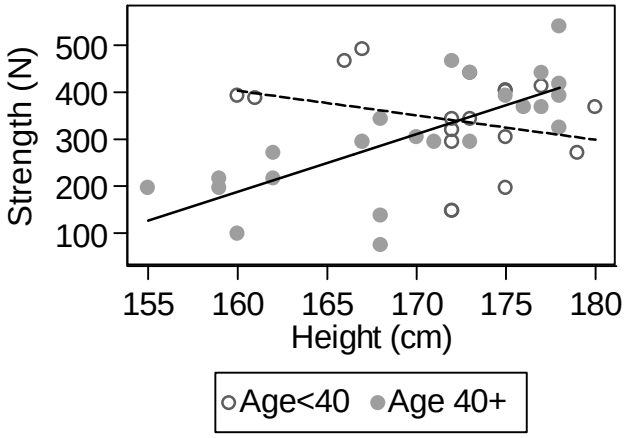


Figure 5. Fitted quadratic curve

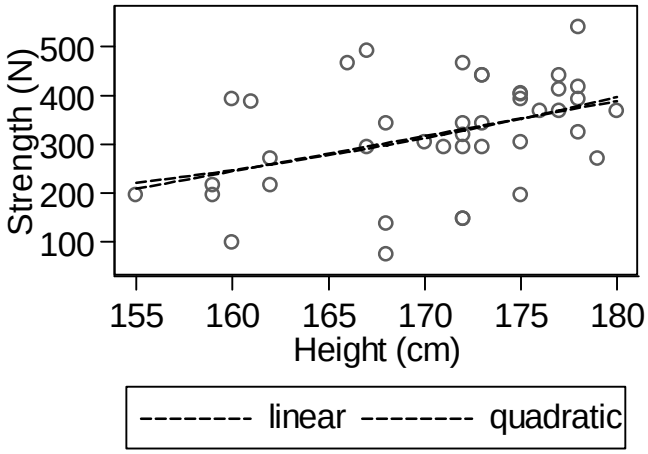


Figure 4 shows this interaction as separate regression lines for younger and older men.

Curvilinear regression

So far, we have assumed that all the regression relationships have been linear, i.e. that we are dealing with straight lines. This is not necessarily so. We may have data where the underlying relationship is a curve rather than a straight line. Unless there is a theoretical reason for supposing that a particular form of the equation, such as logarithmic or exponential, is needed, we test for non-linearity by using a polynomial.

Clearly, if we can fit a relationship of the form

strength = constant + constant × height + constant × age

we can also fit one of the form

strength = constant + constant × height + constant × height^2

to give a quadratic equation, which would produce a curve rather than a straight line. We can continue adding powers of height to give equations which are cubic, quartic, etc., which would produce more complex curves.

For the example data we get

$$\text{Strength} = 1693 - 23.70 \times \text{height} + 0.0918 \times \text{height}^2$$
$$\text{P=0.9} \qquad \text{P=0.8}$$

here is no evidence that the quadratic term improves the prediction of strength. Figure 5 shows the curve, which is hard to distinguish from the straight line.

Height and height squared are highly correlated, which can lead to problems in estimation. To reduce the correlation, we can subtract a number close to mean height from height before squaring. For the data of Table 1, the correlation between height and height squared is 0.9998. This is why the height coefficient has changes and become non-significant. Mean height is 170.7 cm, so 170 is a convenient number to subtract. The correlation between height and height minus 170 squared is –0.44, so the correlation has been reduced, though not eliminated. The regression equation is

$$\text{strength} = -961 + 7.49 \times \text{height} + 0.092 \times (\text{height} - 170)^2$$
$$\text{P=0.01} \qquad \text{P=0.8}$$

The coefficient and P value for the quadratic term have not changed, but the coefficient for the linear term, height, has returned to something like its former value.

Qualitative predictor variables

The predictor variables height and age are quantitative. In the study from which these data come, we also recorded whether or not subjects had cirrhosis of the liver. Cirrhosis was recorded as 'present' or 'absent', so the variable was dichotomous. It is easy to include such variables as predictors in multiple regression. We create a variable which is 0 if the characteristic is absent, 1 if present, and use this in the regression equation just as we did height. The regression coefficient of this dichotomous variable is the difference in the mean of the outcome variable between subjects with the characteristic and subjects without. If the coefficient in this example were negative, it would mean that subjects with cirrhosis were not as strong as subjects without cirrhosis. In the same way, we can use sex as a predictor variable by creating a variable which is 0 for females and 1 for males. The coefficient then represents the difference in mean between male and female. If we use only one, dichotomous predictor variable in the equation, the regression is exactly equivalent to a two sample t test between the groups defined by the variable.

For the strength data, we define a variable cirrhosis = 1 if subject has cirrhosis, 0 if not.

$$\text{Strength} = -544 + 5.86 \times \text{height} - 2.75 \times \text{age} - 34.5 \times \text{cirrhosis}$$
$$\text{P=0.03} \qquad \text{P=0.07} \qquad \text{P=0.3}$$

Men with cirrhosis have mean strength lower than men without cirrhosis, of the same height and age, 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** and the categorical variables (here cirrhosis) are called **factors**. We can also have factors with more than two categories or classes, see below.

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:

$$R^2 = \frac{\text{SS of deviations}}{\text{SS of outcome variable}}$$

Figure 6. Distribution of residuals

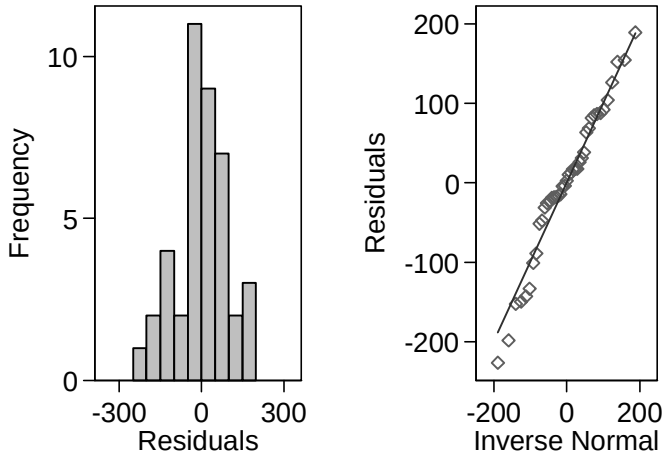
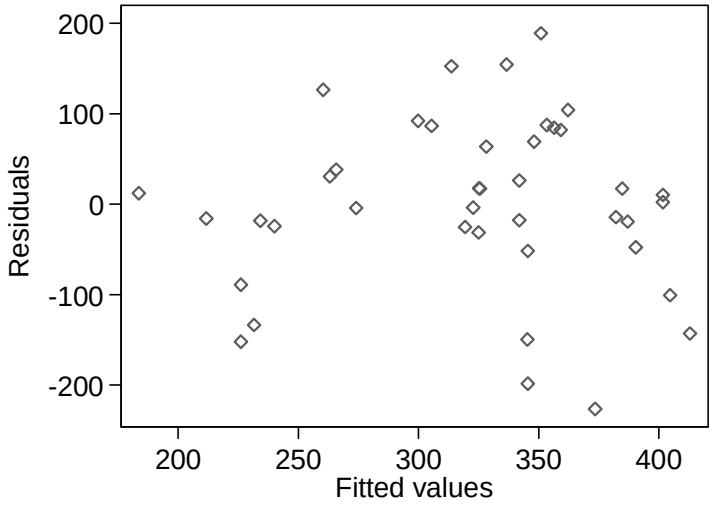


Figure 7. Residuals against predicted or fitted values



R is called the **multiple correlation coefficient**. Unlike the bivariate correlation coefficient, *r*, 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, also called the residuals, follow a Normal distribution with uniform variance. We can check these assumptions by plots of the distribution such as a histogram and a Normal quantile plot of the residuals (Figure 6) and a scatter plot of deviation against predicted values (Figure 7).

Logistic regression

Logistic regression is used when the outcome variable is dichotomous, a ‘yes or no’, whether or not the subject has a particular characteristic such as a symptom. We want a regression equation which will predict the proportion of individuals who have the characteristic, or, equivalently, estimate the probability that an individual will have the characteristic. We cannot use an ordinary linear

regression equation, because this might predict proportions less than zero or greater than one, which would be meaningless. If we used the odds, rather than the proportion, as the outcome we would have a variable which could take any positive value, but could not be negative. We use the log of the odds, also called the **logistic transformation** or **logit** of the proportion as the outcome variable.

The logit can take any value from minus infinity, when the proportion = 0, to plus infinity, when the proportion = 1. We can fit regression models to the logit which are very similar to the ordinary multiple regression models found for data from a Normal distribution. We assume that relationships are linear on the logistic scale. The method is called **logistic regression**, and the calculation is computer intensive. The effects of the predictor variables are found as log odds ratios. We will look at the interpretation in an example.

When giving birth, women who have had a previous caesarean section usually have a trial of scar, that is, they attempt a natural labour with vaginal delivery and only have another caesarean if this is deemed necessary. 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 26.4 Kg/m^2 and for vaginal deliveries the mean was 24.9 Kg/m^2. Two other variables had a strong relationship with a subsequent caesarean. Women who had had a previous vaginal delivery (PVD) were less likely to need a caesarean, odds ratio = 0.18, 95% confidence interval 0.10 to 0.32. Women whose labour was induced had an increased risk of a caesarean, odds ratio = 2.11, 95% confidence interval 1.44 to 3.08. All these relationships were highly significant. The question to be answered was whether the relationship between BMI and caesarean section remained when the effects of induction and previous deliveries were allowed for.

The logistic regression equation predicting the log odds of a caesarean was:

$$\begin{array}{lccccc} \text{log odds caesarean} = -3.70 & + & 0.0883 \times \text{BMI} & + & 0.647 \times \text{induction} & - & 1.80 \times \text{PVD} \\ & & 0.0492 \text{ to } 0.1275 & & 0.228 \text{ to } 1.067 & & -2.38 \text{ to } -1.21 \\ & & P<0.001 & & P=0.003 & & P<0.001 \end{array}$$

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

Because the logistic regression equation predicts the log odds, the coefficients represent the difference between two log odds, a log odds ratio. The antilog of the coefficients is thus an odds ratio. Some programs will print these odds ratios directly. If we antilog the equation we get

$$\begin{array}{lccccccc} \text{odds caesarean} = 0.0247 & \times & 1.092^{\text{BMI}} & \times & 1.910^{\text{induction}} & \times & 0.166^{\text{PVD}} \\ & & 1.050 \text{ to } 1.136 & & 1.256 \text{ to } 2.906 & & 0.09 \text{ to } 0.98 \\ & & P<0.001 & & P=0.003 & & P<0.001 \end{array}$$

This means that induction increases the odds of a caesarean by a factor of 1.910 and a previous vaginal delivery reduces the odds by a factor of 0.166. These are often called **adjusted odds ratios** and 1.91 is the odds ratio for induction of labour adjusted for BMI and previous vaginal delivery. In this example they and their confidence intervals are similar to the unadjusted odds ratios given above, because the three predictor variables happen not to be closely related to each other.

For a continuous predictor variable, such as BMI, the coefficient is the change in log odds for an increase of one unit in the predictor variable. The antilog of the coefficient, the odds ratio, is the factor by which the odds must be multiplied for a unit increase in the predictor. Two units increase in the predictor increases the odds by the square of the odds ratio, and so on. A difference of 5 Kg/m^2 in BMI gives an odds ratio for a caesarean of $1.092^5 = 1.55$, thus the odds of a caesarean are multiplied by 1.55.

Categorical variables with more than two levels

It is straightforward to use qualitative or categorical variables as predictors when there are two groups, but a bit more complicated when there are more. For example, Coste *et al.* (1997) followed up children of short stature given growth hormone treatment. There were three types of treatment: human growth hormone only (311 children), human growth hormone followed by recombinant growth hormone (1455), and recombinant growth hormone only (1467). Hence the treatment is a categorical variable with three categories. If we code these as 1, 2, and 3, then put this variable as a predictor into a multiple or logistic regression, 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. What we do instead is to set up what we call **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 there are categories.

If we put both dummy variables as predictors into a multiple or logistic regression, the coefficient of dummy1 represents the difference between human growth hormone only and recombinant growth hormone only. The coefficient of dummy2 represents the difference between human growth hormone followed by recombinant growth hormone and recombinant growth hormone only. The category represented by all the dummy variables being zero is called the **reference category**, the category to which all the others are compared.

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), and recombinant hormone only 0. The coefficient for recombinant hormone only was 0 by definition, because it was the reference category to which the others were compared. Because it is 0 by definition, it has no confidence interval. The confidence interval for human hormone only is much wider than for the combination treatment, because there are fewer subjects in this category. When we choose the reference category there are two considerations. First, we want a category which gives a meaningful comparison. If there is a control group, we usually choose that. There is no control group here, but we would prefer not to have the more complex regime of both types of hormone. Second, 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.

Treatment regimen was a highly significant predictor of final height, P = 0.0008. There is 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. There is an overall test, called an F test, to test the dummy variables as a group rather than individually.

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. There are some rules of thumb which have been developed using simulation studies:

- for multiple regression, we should have at least 10 observations per variable,
- for logistic regression, we should have at least 10 observations with a 'yes' outcome and 10 observations with a 'no' outcome per variable.

Otherwise, things get very unstable. With logistic regression, we may not be able to trust the large sample P values and confidence intervals.

Types of regression

Multiple regression and logistic regression are the types of regression most often seen in the health care literature in general. There are other types of regression developed for different kinds of outcome variable. These include:

- Cox regression or proportional hazards regression (survival analysis, for data which give the time to an event),
- ordered logistic regression (for outcome variables which are ordered categories, such as health is poor, fair, good, or excellent),
- multinomial regression (for unordered categories, this is very rarely seen),
- Poisson regression (counts, such as the number of deaths per year),
- negative binomial regression (counts with extra variability, such as the numbers of deaths in different case series, becoming increasing popular because there are programs to do it).

J. M. Bland,
1 March 2012

References

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.

Hickish T, Colston K, Bland JM, Maxwell JD. (1989) Vitamin D deficiency and muscle strength in male alcoholics. *Clinical Science* 77, 171-176.