Introduction to Statistics for Clinical Trials

Regression and multiple regression

Martin Bland

Professor of Health Statistics

University of York

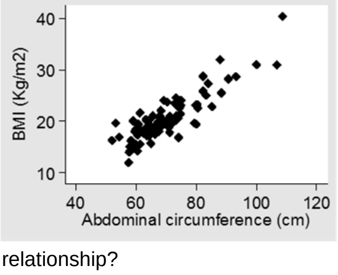
http://www-users.york.ac.uk/~mb55/

Regression analyses

- Simple linear regression
- Multiple linear regression
- Dichotomous predictor variables
- Regression in clinical trials
- Dichotomous outcome variables and logistic regression
- Interactions
- Sample size

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women



(Data of Malcom Savage)

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women.

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

What is the mean BMI for women with any given observed abdominal circumference?

BMI is the **outcome**, **dependent**, **y**, or **left hand side** variable.

Abdominal circumference is the **predictor**, **explanatory**, **independent**, **x**, or **right hand side** variable.

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women.

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

What is the mean BMI for women with any given observed abdominal circumference (AC)?

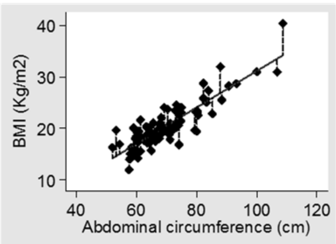
Linear relationship:

$$\text{BMI} = \text{intercept} + \text{slope} \times \text{AC}$$

Equation of a straight line.

Simple Linear Regression

Which straight line should we choose?

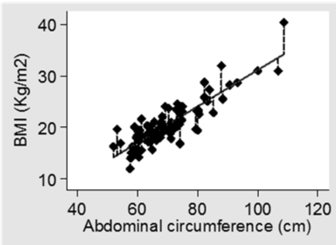


Choose the line which makes the distance from the points to the line ***in the y direction*** a minimum.

Differences between the observed strength and the predicted strength.

Simple Linear Regression

Which straight line should we choose?

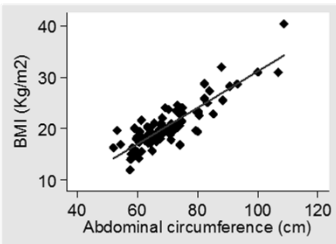


Minimise the sum of the squares of these differences.

Principle of least squares, least squares line or equation.

Simple Linear Regression

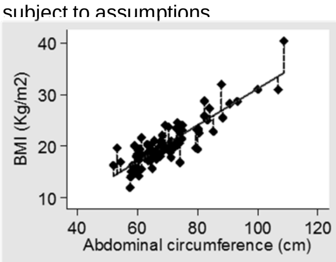
BMI = –4.15 + 0.35 × AC



We can find confidence intervals and P values for the coefficients subject to assumptions.

Simple Linear Regression

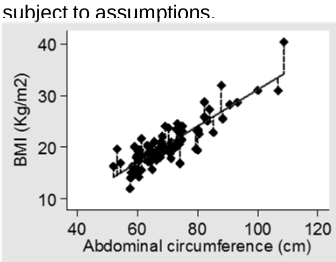
We can find confidence intervals and P values for the coefficients subject to assumptions.



Deviations from line should have a Normal distribution with uniform variance.

Simple Linear Regression

Can find confidence intervals and P values for the coefficients subject to assumptions.



Slope = 0.35 Kg/m²/cm, 95% CI = 0.31 to 0.40 Kg/m²/cm, P<0.001 against zero.

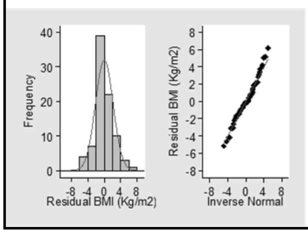
Intercept = –4.15 Kg/m², 95% CI = –7.11 to –1.18 Kg/m².

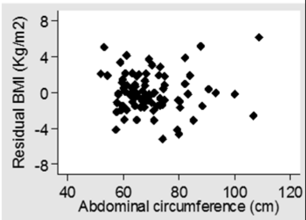
Simple Linear Regression

Assumptions: deviations from line should have a Normal distribution with uniform variance.

Calculate the deviations or residuals, observed minus predicted.

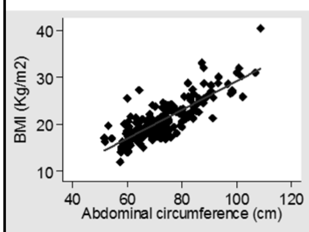
Check Normal distribution: Check uniform variance:

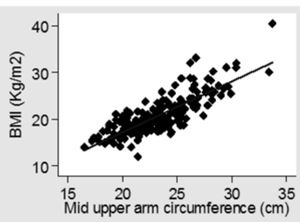




Multiple Linear Regression

More than one predictor:





BMI = –1.35 + 0.31 × AC BMI = –4.59 + 1.09 × MUAC

BMI = –5.94 + 0.18 × AC + 0.59 × MUAC

Multiple Linear Regression

More than one predictor:

BMI = –1.35 + 0.31 × AC   BMI = –4.59 + 1.09 × MUAC

BMI = –5.94 + 0.18 × AC + 0.59 × MUAC

We find the coefficients which make the sum of the squared differences between the observed BMI and that predicted by the regression a minimum.

This is called **ordinary least squares** regression or **OLS** regression.

Multiple Linear Regression

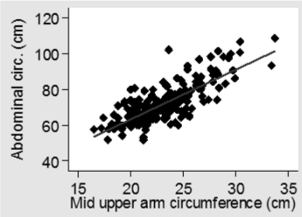
More than one predictor:

BMI = –1.35 + 0.31 × AC   BMI = –4.59 + 1.09 × MUAC

BMI = –5.94 + 0.18 × AC + 0.59 × MUAC

Both coefficients are pulled towards zero because abdominal circumference and arm circumference are related:

MUAC = 7.52 + 2.79 × AC,
P < 0.001



Multiple Linear Regression

More than one predictor:

We can find confidence intervals for the coefficients and test the null hypotheses that coefficients are zero in the population.

BMI =	–5.94	+	0.18 × AC	+	0.59 × MUAC
95% CI	–8.10 to –3.77		0.14 to 0.22		0.45 to 0.74
			P<0.001		P<0.001

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

They can both become not significant, even though the regression as a whole is highly significant.

Multiple Linear Regression

Assumptions:

Just as for simple linear regression, for our confidence intervals and P values to be valid, the data must conform to the assumptions that

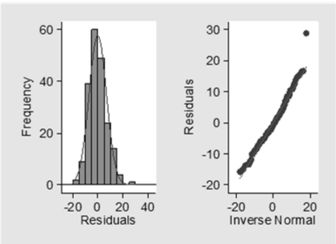
- deviations from line should have a Normal distribution,
- with uniform variance,
- observations must be independent.

Finally, our model of the data is that the relationship with each of our predictors is adequately represented by a straight line rather than a curve.

Multiple Linear Regression

Assumptions:

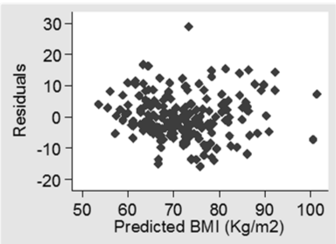
Check by histogram and Normal plot of residuals:



Multiple Linear Regression

Assumptions:

and by plot of residuals against regression estimate:

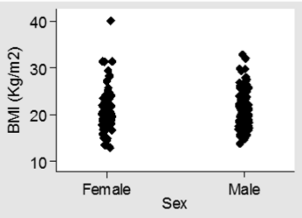


Multiple Linear Regression

Dichotomous predictor: sex.

Variable male = 0 for a female, = 1 for a male.

	BMI =		
	BMI	= 20.51	+ 0.40 × male
95% CI		19.64 to 21.38	–0.75 to 1.55
			P = 0.5



Sex is not a significant predictor alone.

Multiple Linear Regression

Dichotomous predictor: sex.

Variable male = 0 for a female, = 1 for a male.

	BMI =		
	BMI	= 20.51	+ 0.40 × male
95% CI		19.64 to 21.38	–0.75 to 1.55
			P = 0.5

BMI = –6.44	+ 0.18 × AC	+ 0.64 × MUAC	– 1.39 × male
–8.49 to –4.39	0.14 to 0.22	0.50 to 0.78	–1.94 to –0.84
	P<0.001	P<0.001	P<0.001

Male has become a significant predictor because abdominal circumference and arm circumference have removed a lot of variability.

Mean BMI is lower for men than women ***of the same abdominal and arm circumference*** by 1.39 units.

Multiple Linear Regression

Dichotomous predictor: sex.

Variable male = 0 for a female, = 1 for a male.

BMI = –6.44	+ 0.18 × AC	+ 0.64 × MUAC	– 1.39 × male
–8.49 to –4.39	0.14 to 0.22	0.50 to 0.78	–1.94 to –0.84
	P<0.001	P<0.001	P<0.001

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

The continuous variables (here AC and MUAC) are called **covariates**.

The categorical variables (here male sex) are called **factors**.

Regression in clinical trials

Used to adjust for prognostic variables and baseline measurements.

An example: specialist nurse education for acute asthma

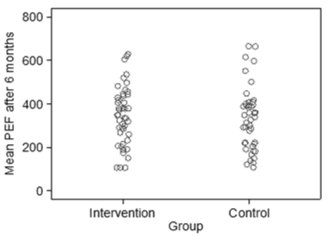
Measurements: peak expiratory flow and symptom diaries made before treatment and after 6 months.

Outcome variables: mean and SD of PEFR, mean symptom score.

Levy ML, Robb M, Allen J, Doherty C, Bland JM, Winter RJD. (2000) A randomized controlled evaluation of specialist nurse education following accident and emergency department attendance for acute asthma. *Respiratory Medicine* 94, 900-908.

Regression in clinical trials

An example: specialist nurse education for acute asthma

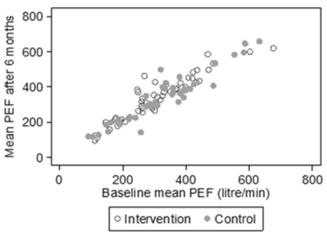


Means: 342 338 litre/min

95% CI (intervention – control) –48 to 63 litre/min, P=0.8.

Regression in clinical trials

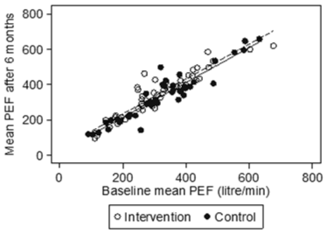
An example: specialist nurse education for acute asthma



If we control for the baseline PEF, we might get a better estimate of the treatment effect because we will remove a lot of variation between people.

Regression in clinical trials

An example: specialist nurse education for acute asthma

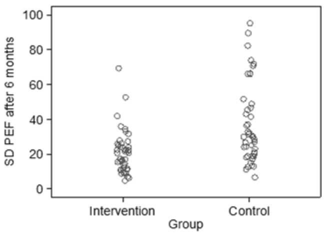


PEF@6m = 58.4 + 0.986 × PEF@base + 20.1 × intervene

	P<0.001	P=0.046
95% CI	0.907 to 1.064	0.4 to 39.7

Regression in clinical trials

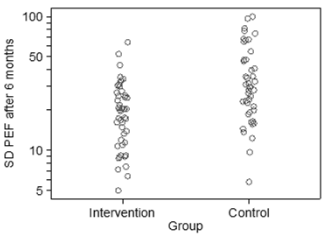
In asthma, large fluctuations in PEF are a bad thing. Use SD.



Clearly we have a skew distribution. Try log transformation.

Regression in clinical trials

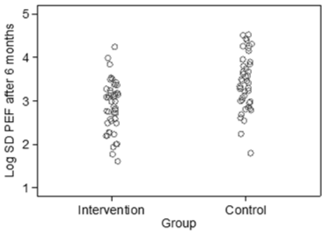
Clearly we have a skew distribution. Try log transformation.



The log scale suggests that this will work.

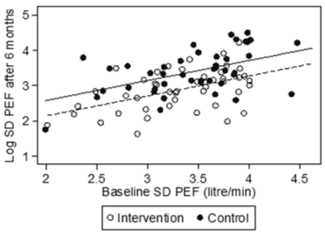
Regression in clinical trials

Log transformed SD of PEF diary, base e.



Regression in clinical trials

Log transformed SD of PEF diary, base e.

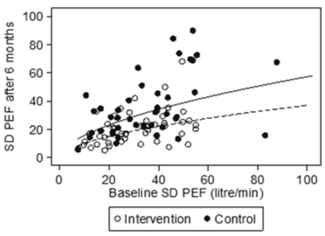


logSD@6m = 0.583 + 0. 564 × logSD@base – 0.435 × intervene

	P<0.001	P<0.001
95% CI	0.356 to 0.771	–0.651 to –0.218

Regression in clinical trials

SD of PEF diary with fitted lines transformed to natural scale.



Estimated treatment effect = –0.435, 95% CI = –0.651 to –0.218.

Back transform, estimated ratio = 0.65, 95% CI = 0.52 to 0.80.

Regression in clinical trials

Advantages

Reduces variability between subjects and so increase power, narrows confidence intervals.

Removes effects of chance imbalances in predicting variables.

Is adjustment cheating?

It can be if we keep adjusting by more and more variables until we have a significant difference.

We should state before we collect the data what we wish to adjust for and stick to it.

Should include any stratification or minimisation variables, centre in multi-centre trials, any baseline measurements of the outcome variable, known important predictors of prognosis.

Dichotomous outcome variables and logistic regression

Factorial clinical trial: Antidepressant drug counselling and information leaflets to improve adherence to drug treatment.

Patients reporting continuing treatment at 12 weeks

Leaflet	Drug counselling		Total
	Yes	No	
Yes	34/52 (65%)	22/53 (42%)	56/105 (53%)
No	32/53 (60%)	20/55 (36%)	52/108 (48%)
Total	66/105 (63%)	42/108 (39%)	

Peveler R, George C, Kinmonth A-L, Campbell M, Thompson C. Effect of antidepressant drug counselling and information leaflets on adherence to drug treatment in primary care: randomised controlled trial. *BMJ* 1999; **319**: 612-615.

Logistic regression

Patients reporting continuing treatment at 12 weeks

Leaflet	Drug counselling		Total
	Yes	No	
Yes	34/52 (65%)	22/53 (42%)	56/105 (53%)
No	32/53 (60%)	20/55 (36%)	52/108 (48%)
Total	66/105 (63%)	42/108 (39%)	

Counselling: P=0.001 Leaflet: P=0.4

Done by logistic regression.

Logistic regression

Patients reporting continuing treatment at 12 weeks

Leaflet	Drug counselling		Total
	Yes	No	
Yes	34/52 (65%)	22/53 (42%)	56/105 (53%)
No	32/53 (60%)	20/55 (36%)	52/108 (48%)
Total	66/105 (63%)	42/108 (39%)	

Our outcome variable is dichotomous, continue treatment yes or no.

We want to predict the proportion who continue treatment.

We would like a regression equation.

Logistic regression

We want to predict the proportion who continue treatment.

We would like a regression equation:

proportion = intercept + slope × counselling + slope × leaflet

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

Predict the log odds of continuing treatment.

log odds = intercept + slope × counselling + slope × leaflet

The slope for counselling will be the increase in the log odds when counselling is used from when counselling is not used.

It will be the log of the odds ratio for counselling, with both the estimate and its standard error adjusted for the presence or absence of the leaflet.

If we antilog, we get the adjusted odds ratio.

Logistic regression

Predict the log odds of continuing treatment.

log odds = intercept + slope × counselling + slope × leaflet

log odds = –0.559 + 0.980 × counselling + 0.216 × leaflet
95% CI 0.426 to 1.53 –0.339 to 0.770
P=0.001 P=0.4

Antilog:

odds = $0.57 \times 2.66^{counselling} \times 1.24^{leaflet}$
95% CI 1.53 to 4.64 0.71 to 2.16

N.B. counselling = 0 or 1, $2.66^0 = 1$, $2.66^1 = 2.66$.

The odds ratio for counselling is 2.66, 95% CI 1.53 to 4.64, P=0.001.

The odds ratio for the leaflet is 1.24, 95% CI 0.71 to 2.16, P=0.4.

Interactions

Does the presence of the leaflet change the effect of counseling?

Define an interaction variable = 1 if we have both counseling and leaflet, zero otherwise.

The counseling and leaflet variables are both 0 or 1.

Multiply the counseling and leaflet variables together.

Interaction = counseling × leaflet.

log odds = intercept + slope × counselling + slope × leaflet + slope × interaction

log odds = –0.560 + 0.981 × counselling + 0.217 × leaflet – 0.002 × interaction
95% CI 0.203 to 1.78 –0.558 to 0.991 –1.111 to 1.107
P=0.01 P=0.6 P=1.0

Interactions

interaction = counseling × leaflet.

log odds = –0.560 + 0.981 × counselling + 0.217 × leaflet – 0.002 × interaction
95% CI 0.203 to 1.78 –0.558 to 0.991 –1.111 to 1.107
P=0.01 P=0.6 P=1.0

Compare the model without the interaction:

log odds = –0.559 + 0.980 × counselling + 0.216 × leaflet
95% CI 0.426 to 1.53 –0.339 to 0.770
P=0.001 P=0.4

The estimates of the treatment effects are unchanged by adding this non-significant interaction but the confidence intervals are wider and P values bigger.

We do not need the interaction in this trial and should omit it.

Interactions

BMI data: interaction between AC and MUAC.

interaction = AC × MUAC

BMI = –6.44 + 0.18 × AC + 0.64 × MUAC – 1.39×male
P<0.001 P<0.001 P<0.001

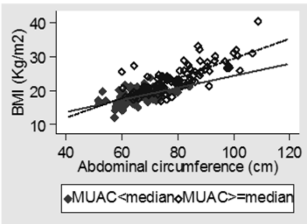
Adding the interaction term:

BMI = 8.45 – 0.02 × AC + 0.03 × MUAC – 1.22 × male + 0.0081 × AC × MUAC
P<0.8 P<0.9 P<0.001 P=0.01

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

Interactions

BMI data: interaction between AC and MUAC.



Interactions

BMI data: interaction between AC and MUAC.

interaction = AC × MUAC

Adding the interaction term:

BMI = 8.45 – 0.02 × AC + 0.03 × MUAC – 1.22 × male + 0.0081 × AC × MUAC
P<0.8 P<0.9 P<0.001 P=0.01

The coefficient for AC now depends on MUAC:

slope = –0.02 + 0.0081 × MUAC

The slope for MUAC depends on AC:

slope = 0.03 + 0.0081 × AC

We cannot interpret the main effects on their own.

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 clinical trials and in the medical literature in general.

There are many other types 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)