The University Of Sheffield.

Critical Numbers 2016, Lecture 6

Regression

Dr Evangelos Kritsotakis
Lecturer in Epidemiology & Statistics
School of Health and Related Research (ScHARR)

e.kritsotakis@sheffield.ac.uk

Learning objectives

- Understand the **role of multivariable regression models** in **controlling confounding** and **prediction**.
- Interpret **scatterplots** for continuous bivariate data in terms of linearity, direction and strength of an association.
- Describe what is meant by a **linear relationship**; understand the concept of the **regression line** and how the **linear regression equation** can be used to model it.
- Be able to correctly interpret the conceptual and practical meaning of **model coefficients**, their confidence intervals and p-values in **linear**, **logistic**, **Poisson** and **Cox regression** analyses.
- **Interpret in context** the results of multivariable regression analyses published in the medical literature.

Overview of regression models

Conceptual framework

Essential work in clinical research pertains to three fundamental subtypes of medical "gnosis":

- **diagnosis** – knowing if disease is present,
- **aetiognosis** (aetiology) – knowing what factors cause the disease,
- **prognosis** – knowing about the future course of a patient's current standing, including how prospects would depend on the choice of intervention or treatment.

Multivariable (multiple) **regression analysis** is a valuable tool for diagnostic, prognostic and aetiognostic research problems.

Important applications of regression (1)

1. Develop a model for prediction of a clinical outcome

- estimate the risk of future outcomes in individuals based on different combinations of clinical and non-clinical characteristics,
- classify individuals as likely to experience the outcome or not,
- develop prediction rules (scoring systems) to direct further diagnostic evaluations, treatments etc

Prediction research includes both **prognostic and diagnostic** studies. Results are **widely used in clinical practice**:

- Apgar score to determine the prognosis of new-borns,
- APACHE and SAPS scores to predict hospital mortality in critically ill patients,
- Prenatal testing to assess the risk that a pregnant woman will give birth to a baby with Down's syndrome. *BMJ 2009;338:b375*

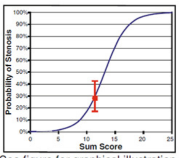
Example (1): Predicting Renal Artery Stenosis

- Diagnostic gold standard is renal angiography (but invasive & costly).
- Can we develop prediction rule for RAS from clinical characteristics, that can be used to select patients for renal angiography?
- **Logistic regression analysis** was performed with data from 477 hypertensive patients who underwent renal angiography.
- Diagnostic accuracy of the regression model was (similar to that of renal scintigraphy): sensitivity = 72% & specificity = 90%.
- It can help to select patients for renal angiography in an efficient manner.

Ann Intern Med 1998;129(9):705-11

Example (1): Predicting Renal Artery Stenosis

	A	B	C	D	E
1	**Prediction rule for renal artery stenosis**				
2					
3	**Predictor**			Value	Score
4	Smoking	former or current =1		1	-
5	Current age	years		45	4.4
6	Gender	male = 1		1	0
7	Atherosclerotic vascular disease*	yes = 1		0	0
8	Onset of hypertension within 2 years	yes = 1		1	1
9	Body mass index >= 25 kg/m2	yes = 1		0	2
10	Presence of abdominal bruit	yes = 1		0	0
11	Serum creatinine concentration	µmol/L		112	4.1
12	Serum cholesterol level > 6.5 mmol/L**	yes = 1		0	0
17	*Sumscore*				*11*
18			Formula		
19	*Predicted probability of renal artery stenosis*		28%		
20	*Confidence interval*		17%	-	43%
21	* femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery				
22	** or cholesterol lowering therapy				



See figure for graphical illustration

Source: E.W. Steyerberg, Clinical Prediction Models

Important applications of regression (2)

2. **Isolate the effect of a single variable on a clinical outcome:**
   - Emphasis on a single effect (e.g. a treatment, an intervention, a risk factor).
   - Need to address this in a multivariable context to **control confounding** (RCTs are not always possible):
   - In a situation of confounding, the crude (unadjusted) data may give us the wrong picture of the effect of the study variable,
   - other variables may be exaggerating the strength of the effect or concealing some or all of it

Example (2): Impact of nosocomial infection on mortality

Factor	No. (%) of patients: Died (n = 111)	No. (%) of patients: Survived (n = 1,721)	Univariate analysis: OR (95% CI)	Univariate analysis: P	Multivariate analysis: aOR (95% CI)	Multivariate analysis: P
Male sex	60 (54.1)	920 (53.5)	1.0 (0.7–1.5)	.903	. . .	. . .
Age ≥65 years	80 (72.1)	876 (50.9)	2.5 (1.6–3.8)	<.001	. . .	. . .
Emergency admission	91 (82.0)	1,141 (66.3)	2.3 (1.4–3.8)	.001	. . .	. . .
Primary admission diagnosis[a]						
Cancer	35 (31.5)	195 (11.3)	3.6 (2.4–5.5)	<.001	. . .	. . .
Respiratory system disease	23 (20.7)	191 (11.1)	2.1 (1.3–3.4)	.003	2.3 (1.3–4.2)	.006
Genitourinary system disease	2 (1.8)	123 (7.1)	0.2 (0.1–1.0)	.046	. . .	. . .
Digestive system disease	7 (6.3)	202 (11.7)	0.5 (0.2–1.1)	.087	. . .	. . .
McCabe-Jackson classification[b]						
Nonfatal disease	34 (30.6)	1,473 (86.3)	Reference	. . .	Reference	. . .
Ultimately fatal disease	52 (46.8)	195 (11.4)	11.6 (7.3–18.3)	<.001	4.9 (2.9–8.3)	<.001
Rapidly fatal disease	25 (22.5)	39 (2.3)	27.8 (15.1–50.9)	<.001	8.7 (4.3–17.6)	<.001
Charlson comorbidity index						
0–1	26 (23.4)	1,116 (64.8)	Reference	. . .	Reference	. . .
2–4	49 (44.1)	465 (27.0)	4.5 (2.8–7.4)	<.001	2.2 (1.3–3.9)	.006
5–12	36 (32.4)	140 (8.1)	11.0 (6.5–18.8)	<.001	2.9 (1.5–5.6)	.001
Karnofsky functional status index[c]						
8–10	16 (14.4)	999 (58.5)	Reference	. . .	Reference	. . .
0–7	95 (85.6)	710 (41.5)	8.4 (4.9–14.3)	<.001	3.2 (1.8–5.7)	<.001
Neutropenia	6 (5.4)	24 (1.4)	4.0 (1.6–10.1)	.003	. . .	. . .
Underwent previous surgical procedure[d]	18 (16.2)	561 (32.6)	0.4 (0.2–0.7)	<.001	. . .	. . .
Exposed to ≥3 invasive devices	31 (27.9)	54 (3.1)	10.6 (6.2–18.0)	<.001	4.2 (2.3–7.5)	<.001
Developed nosocomial infection	36 (32.4)	93 (5.4)	8.4 (5.4–13.2)	<.001	3.6 (2.1–6.1)	<.001

NOTE. OR, odds ratio; aOR, adjusted odds ratio; CI, confidence interval.
[a] Table only includes admission diagnosis categories significantly associated with death in univariate analysis ($P < .20$).
[b] Data missing for 14 patients.
[c] Data missing for 12 patients.
[d] In the 30 days before onset of infection or the 30 days before the survey day.

Infect Control Hosp Epidemiol 2008; 29:685– 692

Example (2): Impact of nosocomial infection on mortality

Univariate analysis does not account for joint effects (looks at one variable at a time). The OR reported is called **crude** or **unadjusted**

The **crude OR** for death is **8.4** (95%CI: 5.4 – 13.2)

Nosocomial infection is associated with a 7.4-fold increase (by 740%) in the odds of death.

But confounding ?

Factor	Died	Survived	OR (95% CI)	P	aOR (95% CI)	P
Developed nosocomial infection	36 (32.4)	93 (5.4)	8.4 (5.4–13.2)	<.001	3.6 (2.1–6.1)	<.001

You know how to calculate the crude OR !

Example (2): Impact of nosocomial infection on mortality

Factor	No. (%) of patients: Died (*n* = 111)	Survived (*n* = 1,721)	Univariate analysis: OR (95% CI)	P	Multivariate analysis: aOR (95% CI)	P
Male sex	60 (54.1)	920 (53.5)	1.0 (0.7–1.5)	.903	. . .	. . .
Age ⩾65 years	80 (72.1)	876 (50.9)	2.5 (1.6–3.8)	<.001	. . .	. . .
Emergency admission	91 (82.0)	1,141 (66.3)	2.3 (1.4–3.8)	.001	. . .	. . .
Primary admission diagnosis[a]						
Cancer	35 (31.5)	195 (11.3)	3.6 (2.4–5.5)	<.001	. . .	. . .
Respiratory system disease	23 (20.7)	191 (11.1)	2.1 (1.3–3.4)	.003	2.3 (1.3–4.2)	.006
Genitourinary system disease	2 (1.8)	123 (7.1)	0.2 (0.1–1.0)	.046		
Digestive system disease	7 (6.3)	202 (11.7)	0.5 (0.2–1.1)	.087		
McCabe-Jackson classification[b]						
Nonfatal disease	34 (30.6)	1,473 (86.3)	Reference	. . .		
Ultimately fatal disease	52 (46.8)	195 (11.4)	11.6 (7.3–18.3)	<.001	[illegible]	
Rapidly fatal disease	25 (22.5)	39 (2.3)	27.8 (15.1–50.9)	<.001	8.[illegible]	
Charlson comorbidity index						
0–1	26 (23.4)	1,116 (64.8)	Reference	. . .	R[illegible]	
2–4	49 (44.1)	465 (27.0)	4.5 (2.8–7.4)	<.001	2.2[illegible]	
5–12	36 (32.4)	140 (8.1)	11.0 (6.5–18.8)	<.001	2.9[illegible]	
Karnofsky functional status index[c]						
8–10	16 (14.4)	999 (58.5)	Reference	. . .	R[illegible]	
0–7	95 (85.6)	710 (41.5)	8.4 (4.9–14.3)	<.001	3.2[illegible]	
Neutropenia	6 (5.4)	24 (1.4)	4.0 (1.6–10.1)	.003		
Underwent previous surgical procedure[d]	18 (16.2)	561 (32.6)	0.4 (0.2–0.7)	<.001		
Exposed to ⩾3 invasive devices	31 (27.9)	54 (3.1)	10.6 (6.2–18.0)	<.001	4.2 (2.3–7.5)	<.001
Developed nosocomial infection	36 (32.4)	93 (5.4)	8.4 (5.4–13.2)	<.001	3.6 (2.1–6.1)	<.001

NOTE. OR, odds ratio; aOR, adjusted odds ratio; CI, confiden…
[a] Table only includes admission diagnosis categories s…
[b] Data missing for 14 patients.
[c] Data missing for 12 patients.
[d] In the 30 days before onset of infection or the 30 da…

Other variables significantly affect mortality, such as age and several baseline factors related to the pre-existing conditions

Notice how imbalanced the distributions are in the two groups !

Example (2): Impact of nosocomial infection on mortality

Multivariate analysis does allow for joint effects. The OR reported is called **adjusted**

The **adjusted OR** for death is **3.6** (95%CI: 2.1 – 6.1)
Nosocomial infection is associated with a 2.6-fold increase (by 260%) in the odds of death, **accounting for** confounding effects by other variables, such as age and severity of pre-existing conditions.

Factor	Died	Survived	OR (95% CI)	P	aOR (95% CI)	P
Developed nosocomial infection	36 (32.4)	93 (5.4)	8.4 (5.4–13.2)	<.001	3.6 (2.1–6.1)	<.001

You need **logistic regression** to calculate the adjusted OR !

Crude vs. Adjusted Effects

- **Crude (or unadjusted)**: does not take into account the effect of confounding variables
- **Adjusted**: accounts for the confounding variable(s)
  Generated using multivariate regression analysis
- Confounding is likely when:

$OR_{crude} \neq OR_{adjusted}$ (logistic regression)
$MD_{crude} \neq MD_{adjusted}$ (linear regression)
$IRR_{crude} \neq IRR_{adjusted}$ (Poisson regression)
$HR_{crude} \neq HR_{adjusted}$ (Cox regression)

Adjusted effects: terminology

The **adjusted OR** is **3.6**
(95%CI: 2.1 – 6.1)

This is an **independent** or **direct effect**
over and above the effects of the other variables.

It was calculated after **accounting (adjusting, correcting, controlling, allowing) for** the effects of other variables in the regression model**.**

There may still be **residual confounding** if we missed important "third" variables in the model
(don't need to worry about this in RCTs)

Important applications of regression (3)

3. Identify multiple independent predictors of a clinical outcome and understand how they jointly affect the outcome

- "independent" in the sense they that have an effect over and above other measured variables.
- need to consider other complexities of how predictors jointly influence the outcome:
  - confounding
  - effect modification (interaction)
  - mediation ("intermediate" variables)

Important applications of regression (4)

4. Covariate adjustment to improve efficiency in RCTs

- The strength of randomization is that comparability is created between the treated groups.
- No systematic confounding can hence occur in RCTs, but random imbalance might occur!
- Some measured baseline variables may be strongly predictive of outcome.

Regression analysis is used to correct for such random imbalances.

Example (4): Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine against Plasmodium falciparum infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet* 2001; 358: 1927–34

Baseline Variables		Vaccine		Control	
Bednet use, n(%)		19	(16%)	10	(9%)
Antibody level, n(%)	Low	48	(38%)	32	(28%)
	Med	38	(30%)	43	(37%)
	High	41	(32%)	40	(35%)
Village, n(%)	BK	40	(31%)	37	(31%)
	BS	15	(11%)	13	(11%)
	HK	12	(9%)	12	(10%)
	KU	12	(9%)	13	(11%)
	SA	28	(21%)	24	(20%)
	TT	24	(18%)	20	(17%)
Age, median (IQR)		25	(20-35)	25	(19-38)

Well balanced distributions of baseline variables.

No confounding issues (as expected!)

Example (4): Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine *Lancet* 2001; 358: 1927–34

Cox regression

	Number developing parasitaemia/total	Crude hazard ratio (95% CI)*	Adjusted hazard ratio (95% CI)†
Group			
RTS,S/AS02	81/131 (62%)	1	1
Control	80/119 (67%)	1·30 (0·95–1·77)‡	1·51 (1·09–2·11)
Village			
Bakadagi	42/77 (55%)	1	1
Bassending	15/28 (54%)	1·47 (0·81–2·65)	1·25 (0·69–2·27)
Hela Kunda	17/24 (71%)	1·45 (0·82–2·54)	1·60 (0·90–2·84)
Kulukuley	23/25 (92%)	2·61 (1·57–4·35)	2·47 (1·44–4·23)
Sanunding	34/52 (65%)	1·64 (1·04–2·57)	2·24 (1·38–3·63)
Touba Tafsir	30/44 (68%)	1·75 (1·10–2·80)	1·87 (1·15–3·06)
Bednet use			
No	141/206 (68%)	1	1
Yes	17/29 (59%)	0·75 (0·45–1·24)	0·93 (0·54–1·58)
Age at enrolment (years)			
18–19	44/60 (73%)	1	1
20–24	42/61 (69%)	0·67 (0·44–1·02)	0·67 (0·43–1·05)
25–26	44/66 (67%)	0·60 (0·40–0·92)	0·70 (0·44–1·11)
37–45	31/63 (49%)	0·34 (0·21–0·54)	0·36 (0·21–0·60)
Concentration of antibody against CSP at enrolment			
<1 mg/L	60/80 (75%)	1	1
1·0–2·7 mg/L	52/81 (64%)	0·68 (0·45–1·01)	0·59 (0·38–0·92)
2·7–42·0 mg/L	43/81 (53%)	0·50 (0·35–0·72)	0·51 (0·34–0·76)

CSP=circumsporozoite protein. *n=250. †n=228. ‡Estimate of crude hazard ratio, using subset of individuals with covariate data, was 1·32.

Example (4): Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine *Lancet* 2001; 358: 1927–34

Cox regression

	Number developing para-sitaemia/total	Crude hazard ratio (95% CI)*	Adjusted hazard ratio (95% CI)†
Group			
RTS,S/AS02	81/131 (62%)	1	1
Control	80/119 (67%)	1·30 (0·95–1·77)‡	1·51 (1·09–2·11)
Village			
Bakadagi	42/77 (55%)		1
Bassending	15/28 (54%)	1·47 (0·81–2·6	1·25 (0·69–2·27)
Hela Kunda	17/24 (7	1·45 (0·82–2·54	60 (0·90–2·84)
Kulukuley	23/	2·61 (1·57–4·35)	1·44–4·23)
		1·64 (1·04–2·57)	8–3·63)
	)	1·75 (1·10–2·80)	06)
		1	
		75 (0·45–1·	
		7 (0·	
		60 (0·	
		34 (0·2	
nst CSP at enro			
	%)	1	
	2/81 (64%)	0·68 (0·45–1·01)	
2·7–42·0 mg/L	43/81 (53%)	0·50 (0·35–0·72)	0·

CSP=circumsporozoite protein. *n=250. †n=228. ‡Estimate of crude hazard ratio, using subset of individuals with covariate data, was 1·32.

In univariate (crude) analysis **HR = 1.30** (risk of malaria is increased by 30% in the non-vaccinated group in our sample.

But the confidence interval includes the null value of 1, so this effect is **not** statistically significant

Example (4): Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine *Lancet* 2001; 358: 1927–34

Cox regression

	Number developing para-sitaemia/total	Crude hazard ratio (95% CI)*	Adjusted hazard ratio (95% CI)†
Group			
RTS,S/AS02	81/131 (62%)	1	
Control	80/119 (67%)	1·30 (0·95–1·77)‡	1· 9–2
Village			
Bakadagi	42/77 (55%)	1	1
Bassending	15/28 (54%)	1·47 (0·81–2·65)	1·25 (0·69
Hela Kunda	17/24 (71%)	1·45 (0·82–2·54)	1·60 (0·90
Kulukuley	23/25 (92%)	2·61 (1·57–4·35)	2·47 (1·4
Sanunding	34/52 (65%)	1·64 (1·04–2·57)	2·24 (1
Touba Tafsir	30/44 (68%)	1·75 (1·10–2·80)	1·87 (
Bednet use			
No	141/206 (68%)	1	1
Yes	17/29 (59%)	0·75 (0·45–1·24)	0·93 (
Age at enrolment (years)			
18–19	44/60 (73%)	1	1
20–24	42/61 (69%)	0·67 (0·44–1·02)	0·67 (
25–26	44/66 (67%)	0·60 (0·40–0·92)	0·70 (
37–45	31/63 (49%)	0·34 (0·21–0·54)	0·36 (
Concentration of antibody against CSP at enrolment			
<1 mg/L	60/80 (75%)	1	1
1·0–2·7 mg/L	52/81 (64%)	0·68 (0·45–1·01)	0·59 (0·38–0·9
2·7–42·0 mg/L	43/81 (53%)	0·50 (0·35–0·72)	0·51 (0·34–0·76)

CSP=circumsporozoite protein. *n=250. †n=228. ‡Estimate of crude hazard ratio, using subset of individuals with covariate data, was 1·32.

In **crude analysis** baseline variables are predictive of malaria risk!

(e.g. increasing age and antibody levels have protective effects)

Should adjusted for these!

Example (4): Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine *Lancet* 2001; 358: 1927–34

Cox regression

	Number developing para-sitaemia/total	Crude hazard ratio (95% CI)*	Adjusted hazard ratio (95% CI)†
Group			
RTS,S/AS02	81/131 (62%)	1	1
Control	80/119 (67%)	1·30 (0·95–1·77)‡	1·51 (1·09–2·11)
Village			
Bakadagi	42/77 (55%)	1	
Bassending	15/28 (54%)	1·47 (0·81–2·65	·25 (0·69–2·27
Hela Kunda	17/24 (71%)	1·45 (0·82	1·60 (0·90–2·84)
Kulukuley	23/25 (92%)	2·61 (	2·47 (1·44–4·23)
Sanunding		)	2·24 (1·38–3·63)
Tou		30)	1·87 (1·15–3·06)
			1
			93 (0·54–1
			(0·43–
			0 (0·44–1
			36 (0·21–0·60)
<1			1
1·0–2·7		(0·45–1·01)	0·59 (0·38–0·92)
2·7–42·0 mg/L	43/81 (53%)	0·50 (0·35–0·72)	0·51 (0·34–0·76)

CSP=circumsporozoite protein. *n=250. †n=228. ‡Estimate of crude hazard ratio, using subset of individuals with covariate data, was 1·32.

After adjustment, **adj. HR = 1.51** (hazard of malaria is increased by 51% in the non-vaccinated group in our sample.

And this **is** statistically significant

Examples recap

Example	Clinical objective	Statistical objective	Regression type
Renal Artery Stenosis	Diagnosis	Prediction	Logistic
Nosocomial Infections	Prognosis	Isolate effect	Logistic
Malaria vaccine trial	Prognosis (treatment prospect)	Covariate adjustment	Cox

Responses & predictors

- Regression relates two kinds of variables:
- **Outcome (or response** or dependent**) variable**: for example
  - Blood pressure
  - 90 day mortality
  - Number of CHD admissions
  - Time to infection
- **Explanatory variables (or predictors** or independent**)**: e.g.
  - age
  - sex
  - severity of illness
  - comorbid conditions
  - treatment type

Common Regression Models

Model	Outcome	What is modelled?	Measure of effect
Linear regression	Continuous	Mean	Mean difference (MD)
Logistic regression	Binary	Log(odds)	Odds ratio (OR)
Poisson regression	Binary (count data)	Log(incidence rate)	Incidence rate ratio (IRR)
Cox regression	Time to event (survival time)	Log(hazard rate)	Hazard ratio (HR)

Simple linear regression

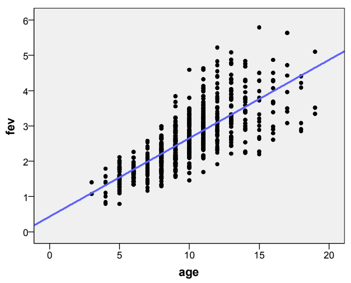
Simple linear regression

- Y = continuous outcome.
- X = explanatory variable (any type)
- “Simple”: only one X variable.
- Aim: Model the dependency of Y on X.
- **How does the mean of Y change with X**?

- E.g. How does **FEV** (=Y) depend on **age** (=X) in children and adolescents?
  - Is there a “linear relationship”? If so,
  - How much increase in FEV do we see, on average, for an increase in age by 1 year?
  - What average FEV would we expect for a given age?

Scatterplot:

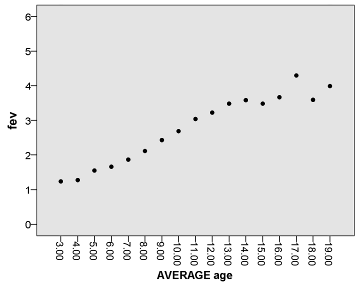
visualizing the relationship between two numerical variables



FEV tends to increase with age, on average.
Points seem to follow a **straight line** with a **positive slope**.
There is a positive **linear relationship (correlation)**.

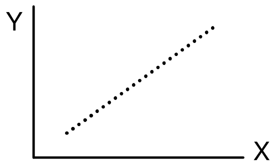
Scatterplot:

visualizing the relationship between two numerical variables



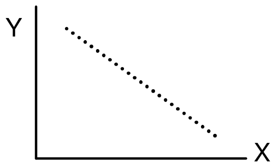
FEV tends to increase with age, on average.
Points seem to follow a **straight line** with a **positive slope**.
There is a positive **linear relationship (correlation)**.

Scatterplots



A perfect positive linear correlation

Y = a + b X
b > 0

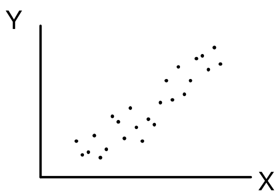


A perfect negative linear correlation

Y = a + b X
b < 0

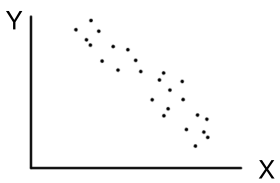
We never see these in real life (variation!)
But we do see linear patterns, linear on average.

Scatterplots



A real-life positive linear correlation

Mean of Y = a + b X
b > 0

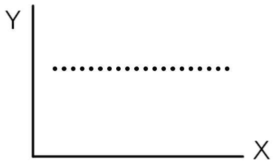


A real-life negative linear correlation

Mean of Y = a + b X
b < 0

Linear patterns, on average.

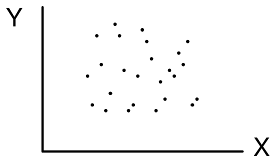
Scatterplots



A perfect **no** relationship

Y = a

b = 0



A real-life **no** relationship

Mean Y = a

b = 0

The **slope coefficient b** tells us if there is a correlation.

Scatterplots

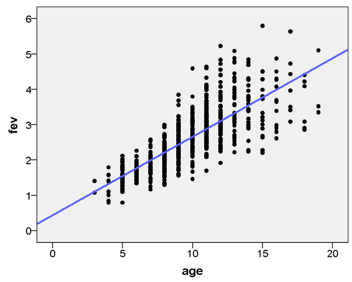
A **non-linear** relationship

Cannot be modelled using a straight line equation

Linear regression cannot handle this directly

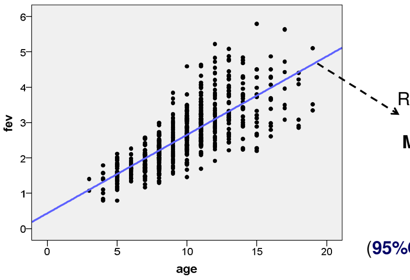
Simple linear regression example

- Data for 654 children and adolescents:
- FEV tends to increase with age, on average: **how can we quantify this effect?**



Critical Numbers

Linear regression: age and FEV in children



Regression line equation:

Mean FEV = a + b · AGE

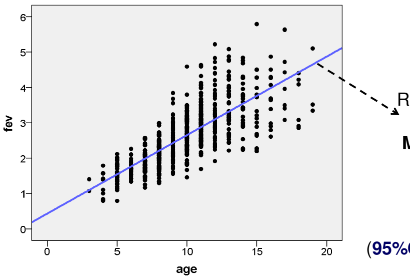
Estimated **b = 0.22**
(sample estimate)

(**95%CI:** 0.21 to 0.24; **p**< 0.001)

The confidence interval excludes 0,
so there is a **statistically significant** linear association.

Critical Numbers

Linear regression: age and FEV in children



Regression line equation:

Mean FEV = a + b · AGE

Estimated **b = 0.22**
(sample estimate)

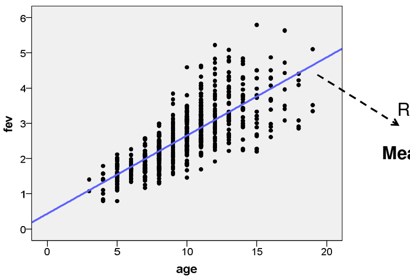
(**95%CI:** 0.21 to 0.24; **p**< 0.001)

The slope coefficient **b = 0.22** quantifies the association:

For each unit increase Age (an increase of 1 year)**,**
we expect an increase of 0.22 litres in FEV, on average.

Critical Numbers

Linear regression: age and FEV in children

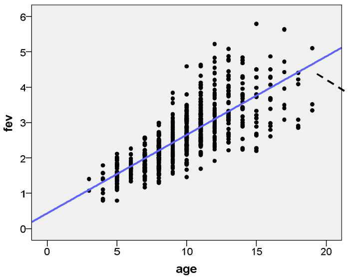


Regression line equation:

Mean FEV = 0.43 + 0.22 AGE

According to this model,
15-year old children are expected to have an average FEV of 3.73 litres
(0.43 + 0.22 x 15).

Linear regression: age and FEV in children



Regression line equation:

Mean FEV = 0.43 + 0.22 AGE

According to this model,
Babies (age = 0 yrs) are expected to have an average FEV = 0.43 litres

(but we should not use this!).

The regression line (the linear regression model)

The regression line can be represented numerically by an equation, which includes two coefficients:

- the ***intercept a*** (the mean value of the outcome, when the predictor variable is equal to zero)
- and the ***slope b*** (the average change in the outcome for a unit change in the x variable):

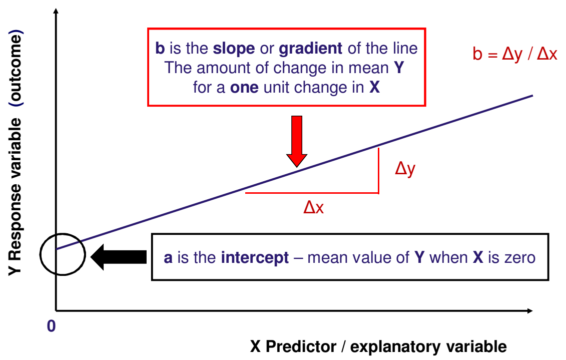
Outcome variable — Predictor variable

Mean y = a + b x

Intercept — Slope

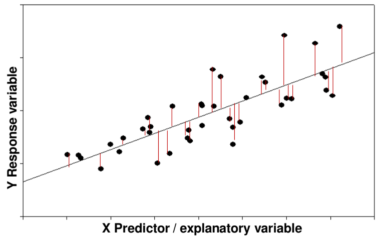
b → average change in y for a unit change in x

Equation of the line: meanY = a + bX



Mathematical estimation of the best fitting line

- The standard way to do this is using a method called **least squares** using a computer.
- The method chooses a line so that the square of the vertical distances between the line and the point (averaged over all points) is minimised.



Uses of linear regression

- **Quantify a linear association**

  e.g. how much increase in FEV we see on average for a year increase in age
- **Predict**

  e.g. what average level of FEV would we expect for a given age, and

  how precise our estimate is for a given age
- **Adjust**

  e.g. what the association between FEV and Age is,

  adjusting for the effect other factors

  such as gender, height and smoking.

Multiple linear regression

Multiple linear regression model

- Simple linear regression model:

mean Y = a + b X

extents to:

- Multiple (multivariable) linear regression model:

$$\textbf{mean } Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$$

Slope coefficients b_i show the strength and direction of association of Y with each of the X_i's.

Regression analysis produces confidence intervals for b_i's and p-values to test the null effect hypotheses H_0: $b_i = 0$

Interpretation of slope coefficients

- Multiple (multivariate) linear regression model:

$$\textbf{mean } Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$$

Slope coefficients b_i quantify the association between Y and each of the X_i's:

Slope b_i = average change (mean difference) in Y per unit increase in $X_{i,}$ adjusted for all other variables in the model

Intercept a = mean Y value when all X_i are zero (usually of no practical meaning)

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938 pregnant women), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smoking (0 = No, 1 = yes)	-113	24	< 0.001
Weight at initial visit (kg)	8	1	< 0.001
Mother's height (cm)	9	2	< 0.001
Age (yrs)	1	21	0.76
Multiparous (0 = No, 1 = yes)	120		< 0.001
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
Educational level (Ref. = University)			
Primary school	-43	31	0.160
Secondary school	-65	25	0.008
Technical school	-50	33	0.130

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smoking (0 = No, 1 = yes)	-113	24	< 0.001
Weight at initial visit (kg)			
Mother's height (cm)			
Age (yrs)			
Multiparous (0 = No, 1 = yes)			
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
Educational level (Ref. = University)			
Primary school	-43	31	0.160
Secondary school	-65	25	0.008
Technical school	-50	33	0.130

Chronic hypertension is of principal focus, but other variables are included since the authors believed that they needed to be adjusted for.

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0			< 0.001
Smoking (0 = No, 1 = ye		24	< 0.001
Weight at initial visit (kg)		1	< 0.001
Mother's height (cm)	9	2	< 0.001
Age (yrs)	1	21	0.76
Multiparous (0 = No, 1 = yes)	120		< 0.001
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
Educational level (Ref. = University)			
Primary school	-43	31	0.160
Secondary school	-65	25	0.008
Technical school	-50	33	0.130

p –values tell us which predictors have no statistically significant effect on birth weight (p > 0.05).

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b	SE	P value
Chronic hyperte	61	48	< 0.001
Smoking (0 = N		24	< 0.001
Weight at initial		1	< 0.001
Mother's heigh		2	< 0.001
Age (yrs)		21	0.76
Multiparous (0 = No, 1 = yes)	120		< 0.001
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
Educational level (Ref. = University)			
Primary school	-43	31	0.160
Secondary school	-65	25	0.008
Technical school	-50	33	0.130

Standard errors can be used to calculate confidence interval for the b coefficients:

$b \pm 1.96$ SE

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smoking (0 = No, 1 = yes)	[illegible]	[illegible]	0.001
Weight at initial visit (kg)	[illegible]	[illegible]	0.001
Mother's height (cm)	[illegible]	[illegible]	0.001
Age (yrs)	[illegible]	[illegible]	.76
Multiparous (0 = No, 1 = yes)	120		< 0.001
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	[illegible]	52	0.007
Other origin	[illegible]	[illegible]	[illegible]
Educational level (Ref. = University)			
Primary school	[illegible]	[illegible]	[illegible]
Secondary school	[illegible]	[illegible]	[illegible]
Technical school	-50	33	0.130

Note the 0 – 1 **coding** for binary variables.
0 is assigned to the reference (control) category

The reference category is explicitly defined for categorical predictors.

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smoking (0 = No, 1 = yes)	-113	24	< 0.001
Weight at initial visit (kg)	8	1	< 0.001
Mother's height (cm)	9	2	< 0.001
Age (yrs)	[illegible]	[illegible]	76
Multiparous (0 = No, 1 = yes)	[illegible]	[illegible]	001
Ethnic group of origin (Ref. = Western Europea			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
Educational level (Ref. = University)			
Primary school	-43	31	0.160
Secondary school	-65	25	0.008
Technical school	-50	33	0.130

Continuous predictors do not have a reference category.

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chr	-161	48	< 0.001
Sm	-113	24	< 0.001
Wei	8	1	< 0.001
Mot	9	2	< 0.001
Age (yrs)	1	21	0.76
Multiparous (0 = No, 1 = yes)	120		< 0.001
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
Educational level (Ref. = University)			
Primary school	-43	31	0.160
Secondary school	-65	25	0.008
Technical school	-50	33	0.130

b coefficients quantify the effect of each predictor on birth weight, adjusting for all the other predictors

Critical Numbers 52

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smoking (0 = No, 1 = yes)	-113	24	< 0.001
Weight at initial visit (kg)	8	1	< 0.001
Mother's height (cm)	9	2	< 0.001
A	1	21	0.76
	120		< 0.001
	108	37	0.004
	-140	52	0.007
	19	33	0.560
	-43	31	0.160
	-65	25	0.008
	-50	33	0.130

b = 9 for mother's height

An increase of 1 cm in mother's height is expected to produce an average increase in birth weight of 9 grams

(10cm → 90 grams)

Not really an impressive effect!

Critical Numbers 53

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smo	-113	24	< 0.001
	8	1	< 0.001
	9	2	< 0.001
	1	21	0.76
	120		< 0.001
	108	37	0.004
	-140	52	0.007
	19	33	0.560
	-43	31	0.160
	-65	25	0.008
	-50	33	0.130

b = -161 for chronic hypertension

An increase of 1 unit in chronic hypertension (from 0=No to 1=Yes) is expected to produce an average decrease in birth weight of 161 grams.

i.e.

Mothers with chronic hypertension have babies with lower birth weights on average;

the absolute mean difference is estimated to be 161 grams (95%CI: 161±1.96x48 → 67 to 255) lower for those mothers

Critical Numbers 54

Example: Effect of chronic hypertension on mean birth weight values (g), multiple linear regression (n = 1,938), France, 1991-1993

Am J Epidemiol 1997;145(8):689-95.

Variable	b coefficient	SE	P value
Chronic hypertension (0 = No, 1 = yes)	-161	48	< 0.001
Smoking (0 = No, 1 = yes)	-113	24	< 0.001
Weight at initial visit (kg)	8	1	< 0.001
Mother's height (cm)	9	2	< 0.001
Age (yrs)	1	21	0.76
Multiparous (0 = No, 1 = yes)	120		< 0.001
Ethnic group of origin (Ref. = Western European)			
North African	108	37	0.004
Sub-Saharan African	-140	52	0.007
Other origin	19	33	0.560
	-43	31	0.160
	-65	25	0.008
	-50	33	0.130

b = -140

How would you interpret this?

Other multiple (multivariable) regression models

Common Regression Models

Y = outcome (response) variable
$X_1, X_2, X_3, \ldots$ = explanatory predictors

Model	Equation
Linear regression	Mean of Y = $a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$
Logistic regression	Log(odds) of Y = $a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$
Poisson regression	Log(incidence rate) of Y = $a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$
Cox regression	Log(hazard rate) of Y = $a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$

All "linear" models!

Logistic regression

Equation
Log(odds) of Y = $a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$

Y = outcome (response) variable, **binary**
$X_1, X_2, X_3, \ldots$ = explanatory predictors

Slope b_i = change in the log odds of Y per unit increase in X_i, adjusted for all other variables in the model.

Logistic regression

Equation

Log(odds) of Y = a + $b_1X_1 + b_2X_2 + b_3X_3$ + …

Exponentiation of slope b_i:

e^{bi} = change in the odds ratio of Y
per unit increase in X_i,
adjusted for all other variables in the model

Impact of nosocomial infection on mortality

Factor	No. (%) of patients: Died (n = 111)	Survived (n = 1,721)	Univariate analysis: OR (95% CI)	P	Multivariate analysis: aOR (95% CI)	P
Male sex	60 (54.1)	920 (53.5)	1.0 (0.7–1.5)	.903	. . .	. . .
Age ⩾65 years	80 (72.1)	876 (50.9)	2.5 (1.6–3.8)	<.001	. . .	. . .
Emergency admission	91 (82.0)	1,141 (66.3)	2.3 (1.4–3.8)	.001	. . .	. . .
Primary admission diagnosis[a]						
Cancer	35 (31.5)	195 (11.3)	3.6 (2.4–5.5)	<.001	. . .	. . .
Respiratory system disease	23 (20.7)	191 (11.1)	2.1 (1.3–3.4)	.003	2.3 (1.3–4.2)	.006
Genitourinary system disease	2 (1.8)	123 (7.1)	0.2 (0.1–1.0)	.046	. . .	. . .
Digestive system disease	7 (6.3)	202 (11.7)	0.5 (0.2–1.1)	.087	. . .	. . .
McCabe-Jackson classification[b]						
Nonfatal disease	34 (30.6)	1,473 (86.3)	Reference	. . .	Reference	. . .
Ultimately fatal disease	52 (46.8)	195 (11.4)	11.6 (7.3–18.3)	<.001	4.9 (2.9–8.3)	<.001
Rapidly fatal disease	25 (22.5)	39 (2.3)	27.8 (15.1–50.9)	<.001	8.7 (4.3–17.6)	<.001
Charlson comorbidity index						
0–1	26 (23.4)	1,116 (64.8)	Reference	. . .	Reference	. . .
2–4	49 (44.1)	465 (27.0)	4.5 (2.8–7.4)	<.001	2.2 (1.3–3.9)	.006
5–12	36 (32.4)	140 (8.1)	11.0 (6.5–18.8)	[illegible]	2.9 (1.5–5.6)	.001
Karnofsky functional status index[c]						
8–10	16 (14.4)	999 (58.5)	[illegible]	. . .	Reference	. . .
0–7	95 (85.6)	710 (4[illegible]	[illegible]–14.3)	<.001	3.2 (1.8–5.7)	<.001
Neutropenia	[illegible]	[illegible]	[illegible]	.003	. . .	. . .
Underwent [illegible]	[illegible]	[illegible]	[illegible]	<.001	. . .	. . .
Exposed to ⩾[illegible]	[illegible]	[illegible]	[illegible]	<.001	4.2 (2.3–7.5)	<.001
Developed n[illegible]	[illegible]	[illegible]	[illegible]	<.001	3.6 (2.1–6.1)	<.001

NOTE. OR [illegible]
[a] Table only i[illegible] lysis (P < .20).
[b] Data missin[illegible]
[c] Data missing[illegible]
[d] In the 30 days before onset of infection or the 30 days before the survey day.

e^b = OR = 2.2 for 2-4 comorbidities
e^b = OR = 2.9 for 5-12 comorbidities
How would you interpret this?

Infect Control Hosp Epidemiol 2008; 29:685– 692

Impact of nosocomial infection on mortality

Factor	No. (%) of patients: Died (n = 111)	Survived (n = 1,721)	Univariate analysis: OR (95% CI)	P	Multivariate analysis: aOR (95% CI)	P
Male sex	60 (54.1)	920 (53.5)	1.0 (0.7–1.5)	.903	. . .	. . .
Age ⩾65 years	80 (72.1)	876 (50.9)	2.5 (1.6–3.8)	[illegible]	. . .	. . .
Emergency admission	91 (82.0)	1,141 (66.3)	[illegible]	[illegible]	[illegible]	. . .
Primary admission diagnosis[a]						
Cancer	35 (31.5)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Respiratory system disease	23 (20.7)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Genitourinary system disease	2 (1.8)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Digestive system disease	7 (6.3)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
McCabe-Jackson classification[b]						
Nonfatal disease	34 (30[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Ultimately fatal disease	52 (46[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Rapidly fatal disease	25 (22[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Charlson comorbidity index						
0–1	26 (23.4[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2–4	49 (44.1)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5–12	36 (32.4)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Karnofsky functional status index[c]						
8–10	16 (14.4)	999 [illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0–7	95 (85.6)	710 (41.5)	[illegible]	[illegible]	[illegible]	<.001
Neutropenia	6 (5.4)	24 (1.4)	4.0 [illegible]	[illegible]	[illegible]	. . .
Underwent previous surgical procedure[d]	18 (16.2)	561 (32.6)	0.4 (0.2–0.7)	<.001	[illegible]	. . .
Exposed to ⩾3 invasive devices	31 (27.9)	54 (3.1)	10.6 (6.2–18.0)	<.001	4.2 [illegible]3–7.5)	<.001
Developed nosocomial infection	36 (32.4)	93 (5.4)	8.4 (5.4–13.2)	<.001	3.6 (2.1–6.1)	<.001

NOTE. OR, odds ratio; aOR, adjusted odds ratio; CI, confidence interval.
[a] Table only includes admission diagnosis categories significantly associated with death in univariate analysis (P < .20).
[b] Data missing for 14 patients.
[c] Data missing for 12 patients.
[d] In the 30 days before onset of infection or the 30 days before the survey day.

The **adjusted OR** for death is **3.6** (95%CI: 2.1 – 6.1)
Nosocomial infection is associated with a 2.6-fold increase (by 260%) in the odds of death, **accounting for** confounding effects by other variables, such as age and severity of pre-existing conditions.

Infect Control Hosp Epidemiol 2008; 29:685– 692

Logistic regression

Equation
Log(odds) of Y = a + b_1X_1 + b_2X_2 + b_3X_3 + …

Intercept a = value of log odds of Y value when all X_i are zero
(may not have any practical meaning)

Logistic regression

Equation
Log(odds) of Y = a + b_1X_1 + b_2X_2 + b_3X_3 + …

If you know the odds you can calculate the probability, so this is actually a probability model.

Equation

$$\text{Probability of Y} = \frac{1}{1+e^{-(a+b_1X_1+b_2X_2+b_3X_3+\cdots)}}$$

Cox regression

Equation
Log(hazard) of Y = a + b_1X_1 + b_2X_2 + b_3X_3 + …

Y = outcome (response) variable, **time to event**
X_1, X_2, X_3, …. = explanatory predictors

Exponentiation of slope b_i
e^{bi} = change in the hazard ratio of Y
per unit increase in X_i,
adjusted for all other variables in the model

Critical Numbers 64

Example: Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine *Lancet* 2001; 358: 1927–34

Cox regression

	Number developing para-sitaemia/total	Crude hazard ratio (95% CI)*	Adjusted hazard ratio (95% CI)†
Group			
RTS,S/AS02	81/131 (62%)	1	1
Control	80/119 (67%)	1·30 (0·95–1·77)‡	1·51 (1·09–2·11)
Village			
Bakadagi	42/77 (55%)	1	1
Bassending	15/28 (54%)	1·47 (0·81–2·65)	1·25 (0·69–2·27)
Hela Kunda	17/24 (71%)	1·45 (0·82–2·54)	1·60 (0·90–2·84)
Kulukuley	23/25 (92%)	2·61 (1·57–4·35)	2·47 (1·44–4·23)
[illegible]	[illegible]	[illegible]–2·57)	2·24 (1·38–3·63)
[illegible]	[illegible]	[illegible]–2·80)	1·87 (1·15–3·06)
[illegible]	[illegible]	[illegible]	1
[illegible]	[illegible]	[illegible]–1·24)	0·93 (0·54–1·58)
[illegible]	[illegible]	[illegible]	1
20–24	[illegible]	0·67 (0·44–1·02)	0·67 (0·43–1·05)
25–26	44/6[illegible]	[illegible]0 (0·40–0·92)	0·70 (0·44–1·11)
37–45	31/63 (49%)	[illegible]0·21–0·54)	0·36 (0·21–0·60)
Concentration of antibody against CSP [illegible]lment			
<1 mg/L	60/80 (75%)	1	1
1·0–2·7 mg/L	52/81 (64%)	0·68 (0·45–1·01)	0·59 (0·38–0·92)
2·7–42·0 mg/L	43/81 (53%)	0·50 (0·35–0·72)	0·51 (0·34–0·76)

CSP=circumsporozoite protein. *n=250. †n=228. ‡Estimate of crude hazard ratio, using subset of individuals with covariate data, was 1·32.

e^b = HR = **0.59** for 1.0 - 2.7 mg/L
e^b = HR = **0.51** for 2.7- 42.0 mg/L
How would you interpret this?

Critical Numbers 65

Example (4): Malaria vaccine trial

Efficacy of RTS,S/AS02 malaria vaccine *Lancet* 2001; 358: 1927–34

Cox regression

	Number developing para-sitaemia/total	Crude hazard ratio (95% CI)*	Adjusted hazard ratio (95% CI)†
Group			
RTS,S/AS02	81/131 (62%)	1	1
Control	80/119 (67%)	1·30 (0·95–1·77)‡	1·51 (1·09–2·11)
Village			
Bakadagi	42/77 (55%)	1	[illegible]
Bassending	15/28 (54%)	1·47 (0·81–2·65	[illegible]25 (0·69–2·27
Hela Kunda	17/24 (71%)	1·45 (0·82[illegible]	1·60 (0·90–2·84)
Kulukuley	23/25 (92%)	2·61 ([illegible]	2·47 (1·44–4·23)
Sanunding	[illegible]	[illegible])	2·24 (1·38–3·63)
Tou[illegible]	[illegible]	[illegible]30)	1·87 (1·15–3·06)
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]3 (0·54–[illegible]
[illegible]	[illegible]	[illegible]	[illegible] (0·43–[illegible]
[illegible]	[illegible]	[illegible]	[illegible]0 (0·44–1[illegible]
[illegible]	[illegible]	[illegible]	[illegible]36 (0·21–0·60)
[illegible]	[illegible]	[illegible]	1
1·0–2·7 [illegible]	[illegible]	[illegible]0·45–1·01)	0·59 (0·38–0·92)
2·7–42·0 mg/L	43/81 (53%)	0·50 (0·35–0·72)	0·51 (0·34–0·76)

CSP=circumsporozoite protein. *n=250. †n=228. ‡Estimate of crude hazard ratio, using subset of individuals with covariate data, was 1·32.

After adjustment,
adj. HR = 1.51
(hazard of malaria is increased by 51% in the non-vaccinated group in our sample.

What does the confidence interval show?

Critical Numbers 66

Poisson regression

Equation

Log(rate) of Y = a + $b_1X_1 + b_2X_2 + b_3X_3$ + …

Y = outcome (response), **incidence count**
X_1, X_2, X_3, …. = explanatory predictors

Exponentiation of slope b_i
e^{bi} = change in the incidence rate ratio of Y per unit increase in X_i, adjusted for all other variables in the model

Recommended Reading:

- Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. **Prognosis and prognostic research: what, why, and how?** BMJ 2009;338:b375. Available at: http://www.bmj.com/content/338/bmj.b375
- Worster A, Fan J, Ismaila A. **Understanding linear and logistic regression analyses.** CJEM 2007;9(2):111-3.Available at: http://cjem-online.ca/v9/n2/p111
- Walters SJ. **What is a Cox model?** Available at http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/cox_model.pdf

Videos

- **Regression Introduction** by Marcello Pagano. Available at: https://youtu.be/0t9m6mLLps8

Further Reading:

- TripepiG , Jager KJ, Dekker FW Zoccali C. **Linear and logistic regression analysis.** Kidney International 2008;73:806–810. Available at: http://www.nature.com/ki/journal/v73/n7/full/5002787a.html
- vanDijk PC, Jager KJ, Zwinderman AH, Zoccali C, Dekker FW. **The analysis of survival data in nephrology: basic concepts and methods of Cox regression**. Kidney International 2008;74(6):705-9. Available at: http://www.nature.com/ki/journal/v74/n6/full/ki2008294a.html
- Campbell MJ, Swinscow TDV. **Statistics at Square One**, 9th Edition, 1997: **chapters 11 and 12**. Available from: http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one