

Linear Models and Effect Magnitudes for Research, Clinical and Practical Applications

Will G Hopkins

Sportscience 14, 49-57, 2010 (sportsci.org/2010/wghlinmod.htm)

Sport and Recreation, AUT University, Auckland 0627, New Zealand. [Email](#). Reviewer: Alan M Batterham, School of Health and Social Care, Teesside University, Middlesbrough TS1 3BA, UK.

Effects are relationships between variables. The magnitude of an effect has an essential role in sample-size estimation, statistical inference, and clinical or practical decisions about utility of the effect. Virtually every effect in research, clinical and practical settings arises from a linear model, an equation in which a dependent variable equals a sum of predictor variables and/or their products. Linear models allow for the effect of one predictor to be adjusted for the effects of other predictors and for the modeling of non-linearity via polynomials. Effects and models used to estimate them depend on the nature of the dependent variable (continuous, count, nominal) and the predictor variables (numeric, nominal). A continuous dependent gives rise to a difference in a mean with a nominal predictor and a slope or correlation with a numeric predictor. Default magnitude thresholds for difference in a mean come from standardization (dividing by the between-subject standard deviation): 0.2, 0.6, 1.2, 2.0 and 4.0 for small, moderate, large, very large and extremely large. The same thresholds apply to a slope, provided the slope is evaluated as the difference for 2 SD of the predictor. Thresholds for correlations are 0.1, 0.3, 0.5, 0.7 and 0.9. Many effects and errors are uniform across the range of the dependent variable when expressed as percents or factors, and these should be estimated via log transformation. Non-uniformity of error arising from repeated measurement or from different subject groups should be addressed via within-subject modeling or mixed modeling, which also provide estimates of individual responses to treatments. Effects on nominal variables and counts are analyzed with various generalized linear models, where the dependent is the log of either the odds of a classification, the hazard (incidence rate) of an event, or the mean count. The effect is estimated initially as a factor representing a ratio between two groups (or per unit or per 2 SD of a numeric predictor) of either odds of a classification, hazards of an event, counts, or count rates. Effects involving common classifications or events can be converted to differences in percent risk and interpreted with magnitude thresholds of 10, 30, 50, 70 and 90; equivalent odds ratios are 1.5, 3.4, 9.0, 32 and 360. Thresholds for common events can also be derived from standardization of log of time to the event. Both sets of thresholds are similar and correspond to hazard ratios of 1.3, 2.3, 4.5, 10 and 100. For counts and rare events, a consideration of proportions attributable to an effect gives rise to ratio thresholds for counts, hazards, risks or odds of 1.1, 1.4, 2.0, 3.3, and 10. Proportional hazards regression is an advanced form of linear modeling for use with events when hazards change with time but their ratio is constant. KEYWORDS: correlation, count ratio, hazard ratio, minimum clinically important difference, odds ratio, relative risk, risk difference, standardization, transformation.

[Reprint pdf](#) · [Reprint doc](#) · [Slideshow](#) · [Reviewer's Commentary](#) · [Updates](#)

After presenting the [Magnitude Matters](#) slideshow recently in several workshops, I realized that it needed more on the role played by linear modeling in estimation of effects. The

additive nature of the linear model is the basis of adjustment for the effects of other factors to get pure or un-confounded effects and to identify potential mediators or mechanisms of an

effect. The additive nature of linear models also explains why we should use the log of the dependent variable to estimate uniform percent or factor effects. A consideration of the error term in a linear model provides further justification for the use of log transformation, along with the use of the unequal-variances t statistic or mixed modeling in analyses where the error term differs between or within subjects. Finally, the analyses for counts and binary dependent variables make little sense without understanding how the underlying linear models require such strange dependent variables as the log of the odds of a classification or the log of the hazard of a time-dependent event. The new [slideshow](#) addresses all these issues and more, using material from the recent [progressive statistics](#) article (Hopkins et al., 2009) and a book chapter on injury statistics (Hopkins, 2009). The slideshow hopefully represents a useful combination of theory and practical advice for anyone who wants to understand and estimate effects in their research.

For more on the way we infer causality, deal with confounders, and account for mechanisms in the relationships between variables, see the slideshow/article on [research designs](#) (Hopkins, 2008). My article and spreadsheets on [understanding stats via simulations](#) (Hopkins, 2007a) are useful for learning more about log transformation, straightforward analyses, and inferential statistics. Follow [this link](#) to a slideshow that details the various approaches to repeated measures and random effects; I presented it at a conference in 2003, but it is still up to date.

When it comes to actual data analysis, you will need extra help with the practicalities of the use of a spreadsheet or stats package. Peruse the [article](#) on comparing two group means and play with the associated [spreadsheet](#) to come to terms with simple comparisons of means and adjustment for a covariate (Hopkins, 2007b). The [article](#) on the various controlled trials and the associated spreadsheets are a little more advanced and also full of useful material (Hopkins, 2006). See my item on [Sad Stats](#) for an overview of some of the stats packages and for a set of files that are useful for SPSS users. If you already have some experience with the SAS package but need specific advice on Proc Mixed or Proc Genmod, [contact me](#).

[Reviewer's Commentary](#)

The [reprint pdf](#) contains this article with a printer-friendly version of the [slideshow](#) (six slides per page).

Update 9 Sept 2010. Slide showing residuals vs predicted for a dependent requiring log transformation. More information on multinomial regression (e.g., for a Likert scale with few items or skewed responses). Other minor improvements.

Update 28 Aug 2010. Odds-ratio thresholds of 1.5, 3.4, 9.0, 32 and 360 now included as an adjunct to proportion-difference thresholds of 10, 30, 50, 70 and 90 percent when modeling and interpreting common time-independent classifications. These odds-ratio thresholds, which I computed directly from the proportion differences centered on 50% (55 vs 45, 65 vs 35, etc.), agree well with a formula devised by Chinn (2000) to convert an odds ratio to a standardized difference in means ($\ln(\text{odds ratio})/1.81$).

Chinn S (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine* 19, 3127-3131

Hopkins WG (2006). [Spreadsheets for analysis of controlled trials, with adjustment for a subject characteristic](#). *Sportscience* 10, 46-50

Hopkins WG (2007a). [Understanding statistics by using spreadsheets to generate and analyze samples](#). *Sportscience* 11, 23-36

Hopkins WG (2007b). [A spreadsheet to compare means in two groups](#). *Sportscience* 11, 22-23

Hopkins WG (2008). [Research designs: choosing and fine-tuning a design for your study](#). *Sportscience* 12, 12-21

Hopkins WG (2009). Statistics in observational studies. In: Verhagen E, van Mechelen W (editors) *Methodology in Sports Injury Research*. OUP: Oxford. 69-81

Hopkins WG, Marshall SW, Batterham AM, Hanin J (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise* 41, 3-12. [Link to PDF](#).

Published July 2010
©2010

Linear Models and Effect Magnitudes for Research, Clinical and Practical Applications

Will G Hopkins

AUT University, Auckland, NZ

Sportscience 14, 49-57, 2010
(sportssci.org/2010/wghlinmod)

- Importance of Effect Magnitudes
- Getting Effects from Models
 - Linear models; adjusting for covariates; interactions; polynomials
- Effects for a continuous dependent
 - Difference between means; "slope"; correlation
 - General linear models: t tests; multiple linear regression; ANOVA...
 - Uniformity of error; log transformation; within-subject and mixed models
- Effects for a nominal or count dependent
 - Risk difference; risk, odds, hazard and count ratios
 - Generalized linear models: Poisson, logistic, log-hazard
 - Proportional-hazards regression

Background: The Rise of Magnitude of Effects

- Research is all about the **effect** of something on something else.
 - The *somethings* are **variables**, such as measures of physical activity, health, training, performance.
 - An effect is a **relationship** between the values of the variables, for example between physical activity and health.
 - We think of an effect as **causal**: more active → more healthy.
 - But it may be only an **association**: more active ↔ more healthy.
 - Effects provide us with evidence for changing our lives.
- The **magnitude** of an effect is important.
 - In clinical or practical settings: could the effect be harmful or beneficial? Is the benefit likely to be small, moderate, large...?
 - In research settings:
 - Effect magnitude determines **sample size**.
 - **Meta-analysis** is all about averaging magnitudes of study-effects.
 - So various research organizations now emphasize magnitude

Getting Effects from Models

- An effect arises from a **dependent** variable and one or more **predictor** (independent) variables.
 - The relationship between the values of the variables is expressed as an equation or **model**.
- Example of **one predictor**: $\text{Strength} = a + b \cdot \text{Age}$
 - This has the same form as the equation of a line, $Y = a + b \cdot X$, hence the term *linear model*.
 - The model is used as if it means: $\text{Strength} \leftarrow a + b \cdot \text{Age}$.
 - If **Age** is in years, the model implies that older subjects are stronger.
 - The magnitude comes from the "**b**" *coefficient* or *parameter*.
 - Real data won't fit this model exactly, so what's the point?
 - Well, it might fit quite well for children or old folks, and if so...
 - We can predict the average strength for a given age.
 - And we can assess how far off the trend a given individual falls.

- Example of **two predictors**: $\text{Strength} = a + b \cdot \text{Age} + c \cdot \text{Size}$

- Additional predictors are sometimes known as **covariates**.
- This model implies that **Age** and **Size** have effects on strength.
- It's still called a linear model (but it's a plane in 3-D).
- Linear models have an incredible property: they allow us to work out the "**pure**" effect of each predictor.
 - By *pure* here I mean the effect of **Age** on **Strength** for subjects of any given **Size**.
 - That is, what is the effect of **Age** if **Size** is **held constant**?
 - That is, yeah, kids get stronger as they get older, but is it just because they're bigger, or does something else happen with age?
 - The *something else* is given by the "**b**": if you hold **Size** constant and change **Age** by one year, **Strength** increases by exactly "**b**".
 - We also refer to the effect of **Age** on **Strength** **adjusted for Size**, **controlled for Size**, or (recently) **conditioned on Size**.
 - Likewise, "**c**" is the effect of one unit increase in **Size** for subjects of any given **Age**.

- With kids, inclusion of **Size** would reduce the effect of **Age**. To that extent, **Size** is a **mechanism** or **mediator** of **Age**.
- But sometimes a covariate is a **confounder** rather than a mediator.
 - Example: **Physical Activity** (predictor) has a strong relationship with **Health** (dependent) in a sample of old folk. **Age** is a confounder of the relationship, because **Age** causes bad health and inactivity.
 - Again, including potential confounders as covariates produces the pure effect of a predictor.
 - Think carefully when interpreting the effect of including a covariate: is the covariate a mechanism or a confounder?
- If you are concerned that the effect of **Age** might differ for subjects of different **Size**, you can add an interaction...
- Example of an **interaction**:

$$\text{Strength} = a + b \cdot \text{Age} + c \cdot \text{Size} + d \cdot \text{Age} \cdot \text{Size}$$
 - This model implies that the effect of **Age** on **Strength** changes with **Size** in some simple proportional manner (and vice versa).
 - It's still known as a linear model.

- You still use this model to adjust the effect of **Age** for the effect of **Size**, but the adjusted effect changes with different values of **Size**.
- Another example of an interaction:

$$\text{Strength} = a + b \cdot \text{Age} + c \cdot \text{Age} \cdot \text{Age} = a + b \cdot \text{Age} + c \cdot \text{Age}^2$$
 - By interacting **Age** with itself, you get a **non-linear effect** of **Age**, here a **quadratic**.
 - If **c** turns out to be negative, this model implies strength rises to a maximum, then comes down again for older subjects.
 - To model something falling to a minimum, **c** would be positive.
 - To model more complex curvature, add $d \cdot \text{Age}^3$, $e \cdot \text{Age}^4$...
 - These are cubics, quartics..., but it's rare to go above a quadratic.
 - These models are also known as **polynomials**.
 - They are all called linear models, even though they model curves.
 - Use the coefficients to get differences between chosen values of the predictor, and values of predictor and dependent at max or min.
 - Complex curvature needs non-linear modeling (see later) or linear modeling with the predictor converted to a **nominal** variable...

- Group, factor, classification or **nominal variables** as predictors:
 - We have been treating **Age** as a number of years, but we could instead use **AgeGroup**, with several **levels**; e.g., child, adult, elderly.
 - Stats packages turn each level into a **dummy variable** with values of 0 and 1, then treat each as a numeric variable. Example:
 - $\text{Strength} = a + b \cdot \text{AgeGroup}$ is treated as $\text{Strength} = a + b_1 \cdot \text{Child} + b_2 \cdot \text{Adult} + b_3 \cdot \text{Elderly}$, where **Child**=1 for children and 0 otherwise, **Adult**=1 for adults and 0 otherwise, and **Elderly**=1 for old folk and 0 otherwise.
 - The model estimates the mean value of the dependent for each level of the predictor: mean strength of children = $a + b_1$.
 - And the difference in strength of adults and children is $b_2 - b_1$.
 - You don't usually have to know about coding of dummies, but you do when using SPSS for some mixed models and controlled trials.
 - Dummy variables can also be very useful for advanced modeling.
 - For simple analyses of differences between group means with t-tests, you don't have to think about models at all!

- Linear models for **controlled trials**
 - For a study of strength training without a control group: $\text{Strength} = a + b \cdot \text{Trial}$, where **Trial** has values pre, post or whatever.
 - $b \cdot \text{Trial}$ is really $b_1 \cdot \text{Pre} + b_2 \cdot \text{Post}$, with **Pre**=1 or 0 and **Post**=1 or 0.
 - The effect of training on mean strength is given by $b_2 - b_1$.
 - For a study with a control group: $\text{Strength} = a + b \cdot \text{Group} \cdot \text{Trial}$, where **Group** has values expt, cont.
 - $b \cdot \text{Group} \cdot \text{Trial}$ is really $b_1 \cdot \text{ContPre} + b_2 \cdot \text{ContPost} + b_3 \cdot \text{ExptPre} + b_4 \cdot \text{ExptPost}$.
 - The changes in the groups are given by $b_2 - b_1$ and $b_4 - b_3$.
 - The net effect of training is given by $(b_4 - b_3) - (b_2 - b_1)$.
 - Stats packages also allow you to specify this model: $\text{Strength} = a + b \cdot \text{Group} + c \cdot \text{Trial} + d \cdot \text{Group} \cdot \text{Trial}$.
 - Group** and **Trial** alone are known as **main effects**.
 - This model is really the same as the interaction-only model.
 - It does allow easy estimation of overall mean differences between groups and mean changes pre to post, but these are useless.

- Or you can model **change scores** between pairs of trials. Example:
 - $\text{Strength} = a + b \cdot \text{Group} \cdot \text{Trial}$, where **b** has four values, is equivalent to $\text{StrengthChange} = a + b \cdot \text{Group}$, where **b** has just two values (expt and cont) and **StrengthChange** is the post-pre change scores.
- You can include **subject characteristics** as covariates to estimate the way they modify the effect of the treatment. Such **modifiers** or **moderators** account for **individual responses** to the treatment.
 - A popular modifier is the **baseline** (pre) score of the dependent: $\text{StrengthChange} = a + b \cdot \text{Group} + c \cdot \text{Group} \cdot \text{StrengthPre}$.
 - Here the two values of **c** estimate the modifying effect of baseline strength on the change in strength in the two groups.
 - And $c_2 - c_1$ is the net modifying effect of baseline on the change.
 - Bonus: a baseline covariate improves precision of estimation when the dependent variable is noisy.
 - Modeling of change scores with a covariate is built into the controlled-trial spreadsheets at SportsScience.

- You can include the **change score** of another variable as a covariate to estimate its role as a **mediator** (i.e., **mechanism**) of the treatment. Example: $\text{StrengthChange} = a + b \cdot \text{Group} + d \cdot \text{MediatorChange}$.
 - d** represents how well the mediator explains the change in strength.
 - $b_2 - b_1$ is the effect of the treatment when **MediatorChange**=0; that is, the effect of the treatment not mediated by the mediator.
- Linear vs **non-linear** models
 - Any dependent equal to a sum of predictors and/or their products is a linear model.
 - Anything else is non-linear, e.g., an exponential effect of Age, to model strength reaching a plateau rather than a maximum.
 - Almost all statistical analyses are based on linear models.
 - And they can be used to adjust for other effects, including estimation of individual responses and mechanisms.
 - Non-linear procedures are available but are more difficult to use.

- ### Specific Linear Models, Effects and Threshold Magnitudes
- These depend on the four **kinds** (or types) of variable.
 - Continuous** (numbers with decimals): mass, distance, time, current; measures derived therefrom, such as force, concentration, volts.
 - Counts**: such as number of injuries in a season.
 - Ordinal**: values are levels with a sense of rank order, such as a 4-pt Likert scale for injury severity (none, mild, moderate, severe).
 - Nominal**: values are levels representing names, such as injured (no, yes), and type of sport (baseball, football, hockey).
 - As **predictors**, the first three can be simplified to **numeric**.
 - If a polynomial is inappropriate, **parse** into 3-5 levels of a nominal.
 - Example: Age becomes AgeGroup (5-14, 15-29, 30-59, 60-79, >79).
 - Values can also be parsed into equal **quantiles** (e.g., quintiles).
 - If an ordinal predictor such as a Likert scale has only 2-4 levels, or if the values are stacked at one end of the scale, analyze the values as levels of a nominal variable.

- As **dependents**, each type of variable needs a different approach. Summary of main effects and models (with **examples**):

Dependent	Predictor	Effect of predictor	Statistical model
continuous Strength	nominal Trial	difference in means	(un)paired t test; (multiple linear) regression; ANOVA; ANCOVA; general linear; mixed linear
continuous Activity	numeric Age	"slope" (difference per unit of predictor); correlation	
nominal InjuredNY	nominal Sex	differences or ratios of proportions, odds, rates, hazards	logistic regression; log-hazard regression; generalized linear;
nominal SelectedNY	numeric Fitness	"slope" (difference or ratio per unit of predictor)	
count Injuries	nominal Sex	ratio of counts	Poisson regression; generalized linear;
count Medals	numeric Cost	"slope" (ratio per unit of predictor)	

Dependent	Predictor	Effect
continuous Strength	nominal Trial	difference or change in means

- The most common effect statistic, for numbers with decimals (continuous variables).
- Difference* when comparing different groups, e.g., patients vs healthy.
- Change* when tracking the same subjects.
- Difference in the changes* in controlled trials.
- The between-subject **standard deviation** provides default thresholds for important differences and changes.
 - You think about the effect (Δmean) in terms of a fraction or multiple of the SD ($\Delta\text{mean}/\text{SD}$).
 - The effect is said to be **standardized**.
 - The smallest important effect is ± 0.20 (± 0.20 of an SD).

- Example: the effect of a treatment on strength

	Trivial effect (0.1x SD)	Very large effect (3.0x SD)
Strength distribution		
- Interpretation of standardized difference or change in means:

	Cohen	Hopkins
trivial	<0.2	<0.2
small	0.2-0.5	0.2-0.6
moderate	0.5-0.8	0.6-1.2
large	>0.8	1.2-2.0
very large	?	2.0-4.0
extremely large	?	>4.0

Complete scale: trivial 0.2 small 0.6 moderate 1.2 large 2.0 very large 4.0 ext. large

- Relationship of standardized effect to difference or change in percentile:

Standardized effect	Percentile change
0.20	50 → 58
0.20	80 → 85
0.20	95 → 97
0.25	50 → 60
1.00	50 → 84
2.00	50 → 98

- Can't define smallest effect for percentiles, because it depends what percentile you are on.
- But it's a good practical measure.
- And easy to generate with Excel, if the data are approx. normal.

Cautions with Standardizing

- Choice of the SD can make a big difference to the effect.
- Use the **baseline (pre) SD**, never the SD of change scores.
- Standardizing works only when the SD comes from a sample **representative of a well-defined population**.
 - The resulting magnitude applies only to that population.
- Beware of authors who show **standard errors of the mean (SEM)** rather than SD.
 - SEM = SD/ $\sqrt{\text{sample size}}$
 - So effects look a lot bigger than they really are.
 - Check the fine print; if authors have shown SEM, do some mental arithmetic to get the real effect.

Other Smallest Differences or Changes in Means

- Single 5- to 7-pt Likert scales: half a step.
- Visual-analog scales scored as 0-10: 1 unit.
- Athletic performance...

Measures of Athletic Performance

- For fitness tests of **team-sport** athletes, use standardization.
- For top **solo** athletes, an enhancement that results in one extra medal per 10 competitions is the smallest important effect.
 - Simulations show this enhancement is achieved with 0.3 of an athlete's typical variability from competition to competition.
 - Example: if the variability is a coefficient of variation of 1%, the smallest important enhancement is 0.3%.
 - Note that in many publications I have mistakenly referred to 0.5 of the variability as the smallest effect.
 - Moderate, large, very large and extremely large effects result in an extra 3, 5, 7 and 9 medals in every 10 competitions.
 - The corresponding enhancements as factors of the variability are:

trivial 0.3	small 0.9	moderate 1.6	large 2.5	very large 4.0	ext. large
-------------	-----------	--------------	-----------	----------------	------------

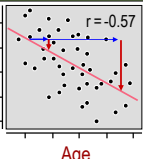
- Beware: smallest effect on athletic performance depends on **method of measurement**, because...
 - A percent change in an athlete's ability to output power results in different percent changes in performance in different tests.
 - These differences are due to the power-duration relationship for performance and the power-speed relationship for different modes of exercise.
 - Example: a 1% change in endurance power output produces the following changes...
 - 1% in running time-trial speed or time;
 - ~0.4% in road-cycling time-trial time;
 - 0.3% in rowing-ergometer time-trial time;
 - ~15% in time to exhaustion in a constant-power test.
 - A hard-to-interpret change in any test following a fatiguing pre-load.

Dependent	Predictor	Effect
continuous Activity	numeric Age	"slope" (difference per unit of predictor); correlation

- A **slope** is more practical than a **correlation**.
- But unit of predictor is **arbitrary**, so it's hard to define smallest effect for a slope.
 - Example: -2% per year may seem trivial, yet -20% per decade may seem large.
 - For consistency with interpretation of correlation, better to express slope as difference per **two SDs** of predictor.
 - It gives the difference between a typically low and high subject.
 - See the page on effect magnitudes at newstats.org for more.
- Easier to interpret the correlation, using **Cohen's scale**.
 - Smallest important correlation is **±0.1**. Complete scale:

trivial	0.1	small	0.3	moderate	0.5	large	0.7	very large	0.9	ext. large
---------	-----	-------	-----	----------	-----	-------	-----	------------	-----	------------

 - But note: in validity studies, correlations >0.90 are desirable.



- The effect of a nominal predictor can also be expressed as a correlation = $\sqrt{\text{fraction of "variance explained"}}$.
 - A 2-level predictor scored as 0 and 1 gives the same correlation.
 - With equal number of subjects in each group, the scales for correlation and standardized difference match up.
 - For >2 levels, the correlation can't be applied to individuals. Avoid.
- Correlations when **controlling for something**...
 - Interpreting slopes and differences in means is no great problem when you have other predictors in the model.
 - Be careful about which SD you use to standardize.
 - But correlations are a challenge.
 - The correlation is either **partial** or **semi-partial** (SPSS: "part").
 - Partial = effect of the predictor within a virtual subgroup of subjects who all have the same values of the other predictors.
 - Semi-partial = unique effect of the predictor with *all* subjects.
 - Partial is probably more appropriate for the individual.
 - Confidence limits may be a problem in some stats packages.

- ### The Names of Linear Models with a Continuous Dependent
- You need to know the jargon so you can use the right procedure in a spreadsheet or stats package.
 - Unpaired t test**: for 2 levels of a single nominal predictor.
 - Use the **unequal-variances** version, never the equal-variances.
 - Paired t test**: as above, but the 2 levels are for the same subjects.
 - Simple linear regression**: a single numeric predictor.
 - Multiple linear regression**: 2 or more numeric predictors.
 - Analysis of variance (ANOVA)**: one or more nominal predictors.
 - Analysis of covariance (ANCOVA)**: one or more nominal and one or more numeric predictors.
 - Repeated-measures analysis of (co)variance**: AN(C)OVA in which each subject has two or more measurements.
 - General linear model (GLM)**: any combination of predictors.
 - In SPSS, nominal predictors are *factors*, numerics are *covariates*.
 - Mixed linear model**: any combination of predictors and errors.

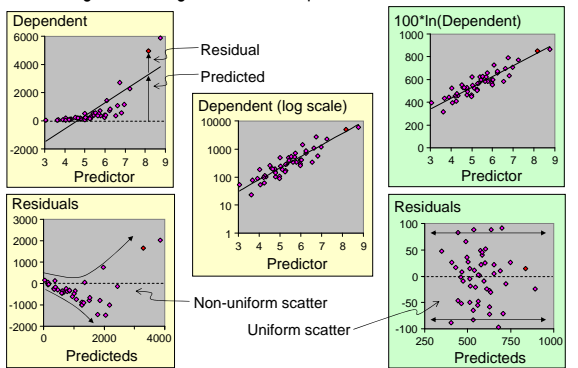
- ### The Error Term in Linear Models with a Continuous Dependent
- Strength = $a + b \cdot \text{Age}$ isn't quite right for real data, because no subject's data fit this equation exactly.
 - What's missing is a different **error** for each subject:

$$\text{Strength} = a + b \cdot \text{Age} + \text{error}$$
 - This error is given an overall mean of zero, and it varies **randomly** (positive and negative) from subject to subject.
 - It's called the **residual error**, and the values are the **residuals**.
 - residual = (observed value) minus (predicted value)
 - In many analyses the error is assumed to have values that come from a **normal** (bell-shaped) distribution.
 - This assumption can be violated a lot. Testing for normality is *not* an issue, thanks to the Central Limit Theorem.
 - With a count as the dependent, the error has a **Poisson** (or the related **negative binomial**) distribution, which *is* an issue.
 - Address with *generalized* linear modeling—see later.

- You characterize the error with a **standard deviation**.
 - It's also known as the standard error of the estimate or the root mean square error.
- In general linear models, the error is assumed to be **uniform**.
 - That is, there is only one SD for the residuals, or the error for every datum is drawn from a single "hat".
 - Non-uniform error** is known as **heteroscedasticity**.
 - If you don't do something about it, you get wrong answers.
- Without special treatment, many datasets show **bigger errors for bigger values** of the dependent.
 - This problem is obvious in some tables of means and SDs, in scatter plots, or in plots of **residual vs predicted values** (see later).
 - Such plots of individual values are also good for spotting outliers.
 - It arises from the fact that effects and errors in the data are **percents** or **factors**, not absolute values.
 - Example: an error or effect of 5% is 5 s in 100 s but 10 s in 200 s.

- Address the problem by analyzing the **log-transformed dependent**.
 - 5% effect means $\text{Post} = \text{Pre} \cdot 1.05$.
 - Therefore $\log(\text{Post}) = \log(\text{Pre}) + \log(1.05)$.
 - That is, the effect is the same for everyone: $\log(1.05)$.
 - And we now have a linear (additive) model, not a non-linear model, so we can use all our usual linear modeling procedures.
- A 5% error means typically $\times/÷ 1.05$, or $\times/÷ 1.05$.
 - And a 100% error means typically $\times/÷ 2.0$ (i.e., values vary typically by a factor of 2), and so on.
- When you finish analyzing the log-transformed dependent, you **back-transform** to a percent or factor effect.
 - Show percents for anything up to ~30%. Show factors otherwise, e.g., when the dependent is a hormone concentration.
- Use the log-transformed values when standardizing.
- Log transformation is often appropriate for a **numeric predictor**.
 - The effect of the predictor is then expressed per percent, per 10%, per 2-fold increase, and so on.

- Example of simple linear regression with a dependent requiring log transformation.
 - A log scale or log transformation produces uniform residuals.



- **Rank transformation** is another way to deal with non-uniformity.
 - You sort all the values of the dependent variable, then rank them (i.e., number them 1, 2, 3,...).
 - You then use this rank in all further analyses.
 - The resulting analyses are sometimes called **non-parametric**.
 - But it's still linear modeling, so it's really parametric.
 - They have names like Wilcoxon and Kruskal-Wallis.
 - Some are truly non-parametric: the sign test; neural-net modeling.
 - Some researchers think you have to use this approach when "the data are not normally distributed".
 - In fact, the rank-transformed dependent is anything but normally distributed: it has a uniform (flat) distribution!!!
 - So it's really an approach to try to get uniformity of effects and error.
 - Problems: it doesn't necessarily give uniformity; you lose a lot of information; it's hard to convert the rank effects back to raw values.
 - So use ranks as a last resort.

- **Non-uniformity** also arises with **different groups and time points**.
 - Example: a simple comparison of means of males and females, with different SD for males and females (even after log transformation).
 - Hence the **unequal-variances t statistic** or test.
 - To include covariates here, you can't use the general linear model: you have to keep the groups separate, as in my spreadsheets.
 - Example: a controlled trial, with different errors at different time points arising from individual responses and changes with time.
 - MANOVA and repeated-measures ANOVA can give wrong answers.
 - Address by reducing or combining repeated measurements into a single change score for each subject: **within-subject modeling**.
 - Then allow for different SD of change scores by analyzing the groups separately, as above.
 - Bonus: you can calculate **individual responses** as an SD.
 - See *Repeated Measures and Random Effects* at sportssci.org and/or the article on the controlled-trial spreadsheets for more.
 - Or specify several errors and much more with a **mixed model**...

- **Mixed modeling** is the cutting-edge approach to the error term.
 - Mixed = fixed effects + random effects.
 - **Fixed effects** are the usual terms in the model; they estimate means.
 - *Fixed*, because they have the same value for everyone in a group or subgroup; they are not sampled randomly.
 - **Random effects** are error terms and anything else randomly chosen from some population; each is summarized with an SD.
 - The general linear model allows only one error. Mixed models allow:
 - specification of different errors between and within subjects;
 - within-subject covariates (GLM allows only subject characteristics or other covariates that do not change between trials);
 - specification of individual responses to treatments and individual differences in subjects' trends;
 - interdependence of errors and other random effects, which arises when you model different lines or curves for each subject.
 - With repeated measurement in controlled trials, simplify analyses by analyzing change scores, even when using mixed modeling.

Dependent	Predictor	Effect
nominal InjuredNY	nominal Sex	differences or ratios of proportions, odds, rates, hazards, mean event time

- For time-dependent effects, subjects start "N" but different proportions end up "Y".
- **Risk or proportion difference = a - b.**
 - Example: $a - b = 83\% - 50\% = 33\%$, so at the time point shown, an extra 33 of every 100 males are injured because they are male.
 - Good for common events, but time-dependent.
 - Can't model risks and estimate differences directly in linear models.
 - Smallest effect: 10% at time when the risk difference is maximum.
 - At that time, 1 male in every 10 is injured due to being male.
 - Complete scale (for common events, where everyone gets affected):
 - trivial 10% small 30% moderate 50% large 70% very large 90% ext. large
 - This scale applies also to time-independent common classifications.

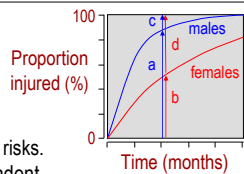
- **Relative risk or risk ratio = a/b.**
 - Example: $83/50 = 1.66$ or "66% increase in risk".
 - Widely used but inappropriate for common time-dependent events.
 - Hazards and hazard ratios are better: see later.
 - For rare events, risk ratio is OK, because same as hazard ratio.
 - Can't estimate directly with linear models.
 - Risk difference or odds ratio are better for common classifications.
 - Magnitude scale: use risk difference, odds ratio or hazard ratio.

For the experts:

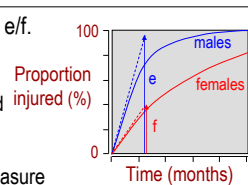
- **Number needed to treat (NNT) = 100/(a - b).**
 - = number you would have to treat or sample for one subject to have an outcome attributable to the effect.
 - Promoted in some clinical journals, but not widely used?
 - Can't estimate directly with linear models.
 - Magnitude scale (if you ever use it) is given by $1/(\text{risk difference})$.

- **Odds ratio** = $(a/c)/(b/d)$.
 - Hard to interpret, but must use to express effects and confidence limits for time-independent classifications, including some case-control designs.
 - Use hazard ratio for time-dependent risks.
 - Magnitudes for common time-independent classifications...
 - *Either* convert to difference in risk between the reference (comparison or control) group and other group. Example shown: if 50% of reference group is affected, and odds ratio is 4.9, then by simple algebra, 83% of other group is affected. Therefore risk difference = 33% (i.e., moderate).
 - *Or* use this scale for odds ratios, which correspond to risk differences of 10, 30, 50, 70 and 90% "centered" on 50%:

trivial	1.5	small	3.4	moderate	9.0	large	32	very large	360	ext. large
---------	-----	-------	-----	----------	-----	-------	----	------------	-----	------------
 - Magnitudes for rare classifications: see later.

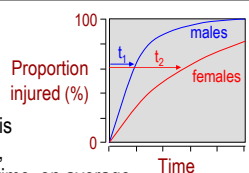


- **Hazard ratio or incidence rate ratio** = e/f .
 - Hazard = instantaneous risk rate = proportion per infinitesimal of time.
 - $e = 100\%/5wk = 20\%/wk = 2.9\%/d$
 - $f = 40\%/5wk = 8\%/wk = 1.1\%/d$
 - $e/f = 100/40 = 20/8 = 2.9/1.1 = 2.5$
 - Hazard ratio is the best statistical measure for time-dependent events.
 - It's the risk ratio *right now*: male risk is 2.5x the female risk.
 - Effects and confidence limits can be derived with linear models.
 - Obviously not dependent on time if hazard rates are constant.
 - And even if both rates change, often OK to assume their ratio is constant, which is the basis of proportional hazards regression.
 - Magnitude scale depends on whether event is common or rare.
 - For common events and constant hazards, maximum risk differences translate into hazard ratios of 1.3, 2.3, 4.4, 10, and 50.
 - Hazard ratios for rare events: see later.



- **Ratio of mean time to event** = t_2/t_1 .
 - Easier for an individual to interpret.
 - If the hazards are constant, it's also the inverse of the hazard ratio.
 - Example: if hazard ratio is 2.5, there is 2.5x the risk of injury. But $1/2.5 = 0.4$, so injury occurs in less than half the time, on average.
- **Difference in mean time to event** = $t_2 - t_1$.
 - Also easy to interpret, but can't model directly.
 - Standardization of the log of individual values of time to event leads to another scale for hazard ratios or mean-time ratios of common events: 1.3, 2.2, 4.5, 13, 100.
 - This scale is similar to that given by consideration of maximum risk difference for common events. Averaging the two and simplifying...
- **Hazard-ratio thresholds** for common events:

trivial	1.3	small	2.3	moderate	4.5	large	10	very large	100	ext. large
---------	-----	-------	-----	----------	-----	-------	----	------------	-----	------------



Magnitude Thresholds for Rare Events and Classifications

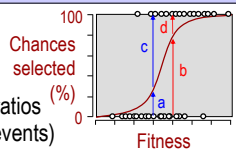
- The focus is the affected few and/or those who deal with them.
- Hazard ratio = risk ratio = odds ratio for low risks (or short times).
- A ratio of 1.1 would produce a 10% increase or decrease in the workload of anyone dealing with the event.
 - Anything less might go unnoticed, so 1.1 is the smallest effect.
- Or in a group of affected individuals, 1 in 10 able to blame the effect represents a defensible smallest effect. Similarly 3, 5, 7 and 9 individuals in 10 represent the other magnitude thresholds.
 - Corresponding hazard ratios are $10/9$, $10/7$, $10/5$, $10/3$, and $10/1$. Hence...
- **Hazard-ratio thresholds** for rare events:

trivial	1.1	small	1.4	moderate	2.0	large	3.3	very large	10	ext. large
---------	-----	-------	-----	----------	-----	-------	-----	------------	----	------------

 - By a similar argument, this scale applies to **count ratios**.
 - Oddly, these thresholds are smaller than those for common events.
 - We tolerate relatively higher risk for a rare event, but if we end up as an event, we wish the risk had been lower!

Dependent	Predictor	Effect
nominal SelectedNY	numeric Fitness	"slope" (difference or ratio per unit of predictor)

- Derive and interpret the "slope" (a correlation isn't defined here).
- As with a nominal predictor, you have to express effects as odds or hazard ratios (for time-independent or -dependent events) to get confidence limits.
 - Example shows how chances would change with fitness, and the meaning of the odds ratio per unit of fitness: $(b/d)/(a/c)$.
 - Odds ratio here is $\sim (75/25)/(25/75) = 9.0$ per unit of fitness.
 - Best to express as odds or hazard ratio per 2 SD of predictor.
 - Magnitude scales are then the same as for nominal predictors.



Dependent	Predictor	Effect
count Injuries	nominal Sex	ratio of counts
count Tackles	numeric Fitness	"slope" (ratio per unit of predictor)

- Effect of a nominal predictor is expressed as a ratio (factor) or percent difference.
 - Example: in their sporting careers, women get 2.3 times more tendon injuries than men.
 - If the ratio is ~ 1.5 or less, it can be expressed as a percent: men get 26% (1.26 times) more muscle sprains than women.
- Effects of a numeric predictor are expressed as factors or percents per unit or per 2 SD of the predictor.
 - Example: 13% more tackles per 2 SD of repeated-sprint speed.
- Magnitude scale for count ratios is same as for rare events:

trivial	1.1	small	1.4	moderate	2.0	large	3.3	very large	10	ext. large
---------	-----	-------	-----	----------	-----	-------	-----	------------	----	------------

Details of Linear Models for Events, Classifications, Counts

- Counts, and binary variables representing levels of a nominal, give wrong answers as dependents in the general linear model.
 - It can predict negative or non-integral values, which are impossible.
 - Non-uniformity would also be an issue.
- Generalized linear modeling has been devised for such variables.
 - The generalized linear model predicts a dependent that *can* range continuously from $-\infty$ to $+\infty$, just as in the general linear model.
 - You specify the dependent by specifying the **distribution** of the dependent and a **link function**.
 - For a continuous dependent, specifying the **normal** distribution and the **identity** link produces the general linear model.
 - Don't use this approach with continuous dependents, because the standard procedures for general linear modeling are easier.
 - Easiest to understand the approach with counts first...

- For **counts** (e.g., each athlete's number of injuries), the dependent is the **log of the mean count**.
 - The mean count ranges continuously from 0 to $+\infty$.
 - The log of the mean count ranges from $-\infty$ to $+\infty$.
 - So the link function is the **log**.
 - Specify the distribution for counts, **Poisson**.
 - The model is called **Poisson regression**.
 - The log link results in effects expressed as count ratios.
 - If the counts accumulate over different periods for different subjects, you can specify the period in the model as an **offset** or denominator.
 - You are then modeling rates, and the effects are rate ratios.
 - Specify a **negative binomial** distribution if you think the events for each subject tend to occur in clusters rather than truly randomly.
 - The model thereby reflects the fact that the counts have a bigger variation for a given predicted count than purely Poisson counts.

- For binary variables representing **time-independent events** (e.g., a classification, such as selected or not), the dependent is the **log of the odds** of the event occurring.
 - Odds= $p/(1-p)$, where p is the probability of the event.
 - P ranges from 0 to 1, so odds range continuously from 0 to $+\infty$.
 - So log of the odds ranges from $-\infty$ to $+\infty$.
 - So the link function is the **log-odds**, also known as the **logit**.
 - Specify the distribution for binary events, **binomial**.
 - The model is called **logistic regression**, but **log-odds regression** would be better.
 - The log of the odds results in effects expressed as odds ratios.
 - A log-odds model may be simplistic or unrealistic, but it's got to be better than modeling p or log p, which definitely does not work.
 - Some researchers mistakenly use this model for time-dependent events, such as development of injury. But...
 - If proportions of subjects experiencing the event are low, you can model risk, odds or hazards, because the ratios are the same.

- For binary variables representing **time-dependent events** (e.g., un/injured), the dependent is the **log of the hazard**.
 - The hazard is the probability of the event per unit time.
 - For events that accumulate with a constant hazard (h), the proportion of subjects affected at time t is given by $p = 1 - e^{-ht}$; hence $h = -\log(1 - p)$.
 - The hazard ranges continuously from 0 to $+\infty$.
 - Log of the hazard ranges from $-\infty$ to $+\infty$.
 - The link function is known confusingly as the **complementary log-log**: $\log(-\log(1-p))$.
 - I prefer to refer to the **log-hazard** link.
 - Specify the distribution for binary events, **binomial**.
 - The model has no common name. I call it **log-hazard regression**.
 - The log of the hazard results in effects expressed as hazard ratios.
 - You can specify a different monitoring time for each subject.
 - When hazards aren't constant, use **proportional hazards regression**.

Three slides for the experts

Other Models for Events and Classifications

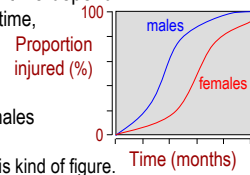
- All have outcomes modeled as ratios (between levels of nominal predictors) or ratios per unit (or per 2 SD) of numeric predictors.
- The magnitude scales for common and rare events and classifications are the same as in previous models.
- Summary (with **examples**):

Dependent	Effect of predictor	Statistical model
multiple proportions choice of several sports	odds ratio	multinomial logistic regression; generalized linear
ordinal injury severity (4-pt Likert)	odds or hazard ratio	cumulative logistic or hazard regression; generalized linear
time to event time to injury	hazard ratio	proportional hazards (Cox) regression

- When the dependent is a **nominal with >2 levels**, group into various combinations of 2 levels and use the above models, or...
- For the advanced class...
 - **Multinomial logistic regression**, for time-independent nominals (e.g., a study of predictors of choice of sport).
 - Use the **multinomial** distribution and the **generalized logit** link (available in SAS in the new Glimmix procedure).
 - SAS does not provide a link in Glimmix or Genmod for multinomial hazard regression of time-dependent nominals.
 - **Cumulative logistic regression**, for time-independent ordinals (e.g., injury severity on a 4-point Likert scale).
 - **Multinomial** distribution; **cumulative logit** link.
 - Use for <5-pt or skewed Likert scales; otherwise use general linear.
 - **Cumulative hazard regression**, for time-dependent ordinals (e.g., uninjured, mild injury, moderate injury, severe injury).
 - **Multinomial** distribution; **cumulative complementary log-log** link.
- Generalized linear models for repeated or clustered measures are also known as **generalized estimating equations**.

- **Proportional hazards (Cox) regression** is another and more advanced form of linear modeling for time-dependent events.

- Use when hazards can change with time, if you can assume ratios of the hazards of the effects are constant.
- Example: hazard changes as the season progresses, but hazard for males is always 1.5x that for females.
 - A constant ratio is not obvious in this kind of figure.
- Time to the event is the dependent, but effects are estimated and interpreted as hazard ratios.
- The model takes account of **censoring**: when someone leaves the study (or the study stops) before the event has occurred.



- Not covered in this presentation: magnitude thresholds for measures of reliability, validity, and diagnostic accuracy.

Main Points

- An effect is a relationship between a dependent and predictor.
- Effect magnitudes have key roles in research and practice.
- Magnitudes are provided by linear models, which allow for adjustment, interactions, and polynomial curvature.
- Continuous dependents need various general linear models.
 - Examples: t tests, multiple linear regression, ANOVA...
 - Within-subject and mixed modeling allow for non-uniformity of error arising from different errors with different groups or time points.
- Effects for continuous dependents are mean differences, slopes (expressed as 2 SD of the predictor), and correlations.
 - Thresholds for small, moderate, large, very large and extremely large standardized mean differences: 0.2, 0.6, 1.2, 2.0, 4.0.
 - Thresholds for correlations: 0.1, 0.3, 0.5, 0.7, 0.9.
 - Many dependent variables need log transformation before analysis to express effects and errors as uniform percents or factors.

- Counts and nominal dependents (representing classifications and time-dependent events) need various generalized linear models.
 - Examples: Poisson regression for counts, logistic regression for classifications, log-hazard regression for events.
 - The dependent variable is the log of the mean count, the log of the odds of classification, or the log of the hazard (instantaneous risk) of the event.
- Effect-magnitude thresholds for counts and nominal dependents:
 - Percent risk differences for classifications: 10, 30, 50, 70, 90.
 - Corresponding odds ratios for classifications: 1.5, 3.4, 9.0, 32, 360.
 - Hazard-ratio thresholds for common events: 1.3, 2.3, 4.5, 10, 100.
 - Ratio thresholds for counts and rare events: 1.1, 1.4, 2.0, 3.3, 10 (apply equally to count, hazard, risk or odds ratios).
- Use proportional hazards regression when hazards vary with time but hazard ratio is constant.

This presentation was downloaded from:

SPORTSCIENCE sportsci.org

A Peer-Reviewed Site for Sport Research

[See Sportsmedicine 14, 2010](#)