Overview of the content

Research Methods for Human Inquiry
Andrew Perfors

We've gone over a lot of stuff…

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
- Regression
- Advanced topics

Contents of RMHI

- Introduction to R
  - Packages, R Markdown, variables, data frames
  - Functions, arguments
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
- Regression
- Advanced topics

Contents of RMHI

- Introduction to R
- Data manipulation
  - Pipes, grouping, summary/descriptive statistics
  - Long vs wide form, filter, mutate
- Figures
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
- Regression

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
  - Scatterplots, boxplots, violin plots, bar plots, histograms (all assessable, but especially these)
  - Geoms, arguments, layering, colour palettes
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
- Regression

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
  - Null hypothesis significance testing; p-values
  - Type 1 and Type 2 error
  - Binomial and normal distributions
  - Sampling distributions; confidence intervals
  - Bayesian vs frequentist approaches
- Chi-square tests
- t-tests
- ANOVA
- Regression

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
  - Goodness of fit test, test of independence
  - Chi-squared test statistic
  - Fisher exact test, McNemar test
  - Cramers' V
- t-tests
- ANOVA
- Regression

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
- t-tests
  - one-sample, paired and independent t-tests
  - t statistic
  - Student's vs Welch's t-tests
  - Cohen's d
  - Shapiro-Wilk test, QQ Plots, Wilcoxon
- ANOVA
- Regression

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
  - one-way ANOVA, two-way ANOVA, interactions
  - post hoc tests and multiple comparison correction
  - effect size with eta-squared
  - F statistic, SSb, SSw
  - Levene test, Welch one-way ANOVA, Kruskal-Wallis
- Regression

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
- Regression
  - Spearman & Pearson correlations
  - Interpretation of regression line; significance; residuals
  - Effect size with R-squared
  - Coefficients and standardised coefficients

Contents of RMHI

- Introduction to R
- Data manipulation
- Figures
- Statistical theory
- Chi-square tests
- t-tests
- ANOVA
- Regression
- Advanced topics
  - Assumptions of linear regression: linearity, normality of residuals, collinearity, high-influence points
  - Model selection, penalising complexity, AIC, BIC

Some things I didn't get to…

This is not examinable! I just want you to know it *exists* and where to find resources for it when/if it comes up for you in the future

- Everything is a linear model and model selection!
- Unbalanced ANOVAs
- Beyond linear regression
- Mixed-effect models
- Bayesian data analysis

Everything is linear model and model selection!

	Outcome variable	Predictor variable(s)
Chi-squared test	counts/frequencies	Categorical
t-test	numeric	1 to 2 groups (categorical)
ANOVA	numeric	3+ groups, or multiple factors
regression	numeric	Numeric

Normal regression $\longrightarrow$ $Y_i = b_1 X_{1i} + b_2 X_{2i} + b_0 + \varepsilon_i$

Equivalent to ANOVA/t-test when slopes b are binary (0 or 1). i.e. group A is 0 and group B is 1

Equivalent to chi-squared when it is log-linear (i.e., take the log of everything)

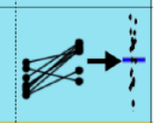
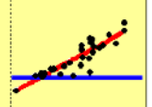
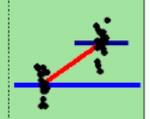
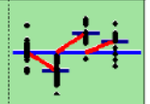
Everything is linear model and model selection!

See Chapter 16 in *Learning Statistics with R* and the linear models cheat sheet linked to in the LMS this week if you want to learn more

Common statistical tests are linear models

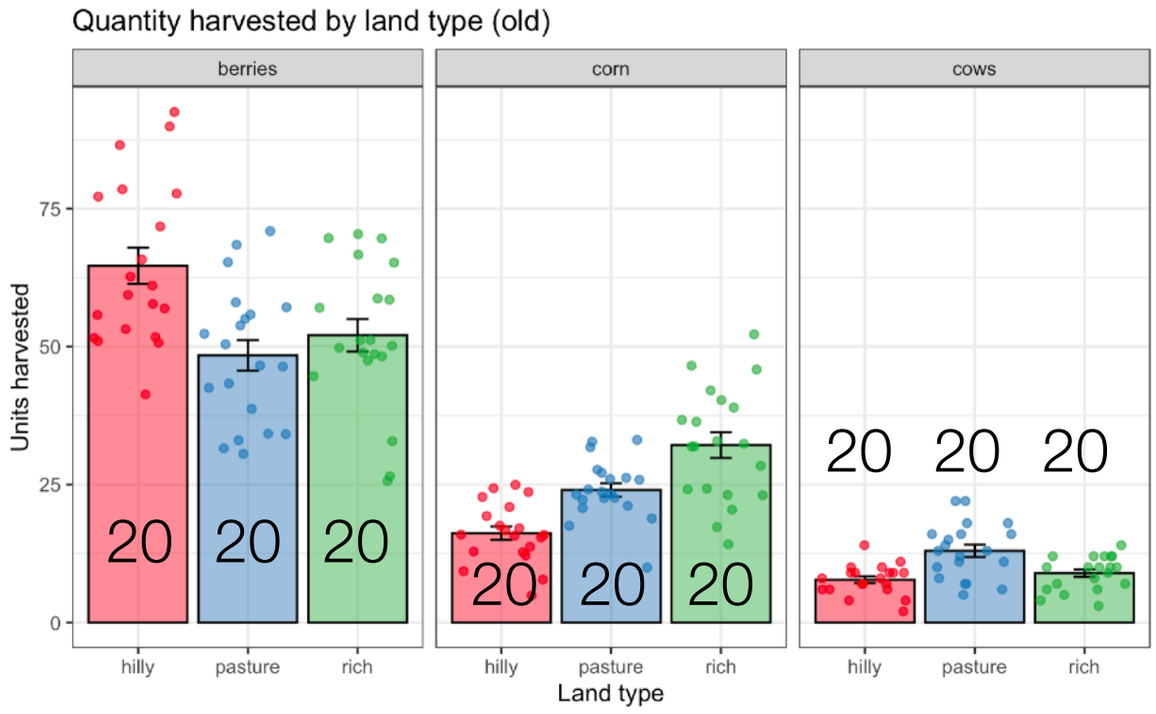
Last updated: 28 June, 2019.Also check out the Python version!

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + S_2 + S_3 +…+ S_K + G_2*S_2 + G_3*S_3 + … + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y** ~ **group** parameters. *Note: $\mathbf{G_{2\,to\,N}}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $\mathbf{S_{2\,to\,K}}$ for sex. The first line (with G_i) is main effect of group, the second (with S_j) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2 + G_3*S_3 +…+ G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_{ij}) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

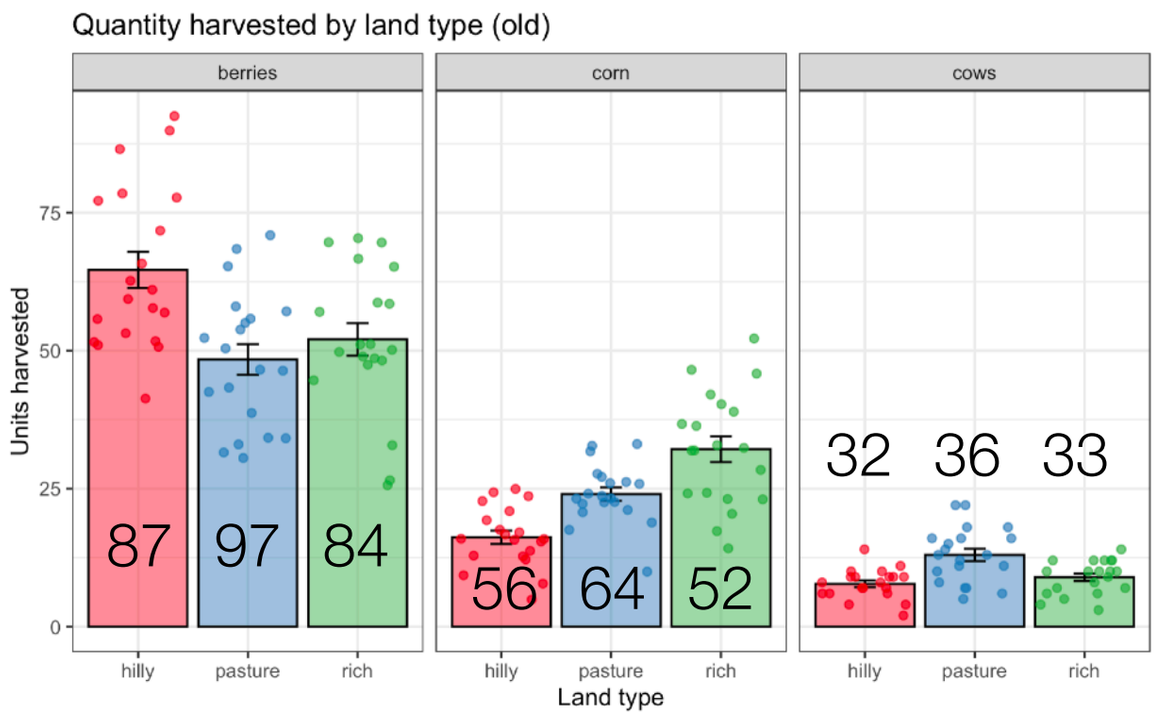
Unbalanced ANOVAs

We covered balanced ANOVAs, which is what happens when there are equal numbers of data points in each group and sub-group



Unbalanced ANOVAs

It's very common to have unbalanced ones, which are more complicated to calculate because the *order* you evaluate variables changes the results



There are three different ways to do things: Type I, Type II, and Type III sums of squares. They work by doing model selection on models with or without each variable, and vary in the order they compare models

Unbalanced ANOVAs

Short version: I recommend Type II sums of squares. Type III is very common. To run them you use the `lm()` function followed by `Anova()` from the `car` package, where you specify the type

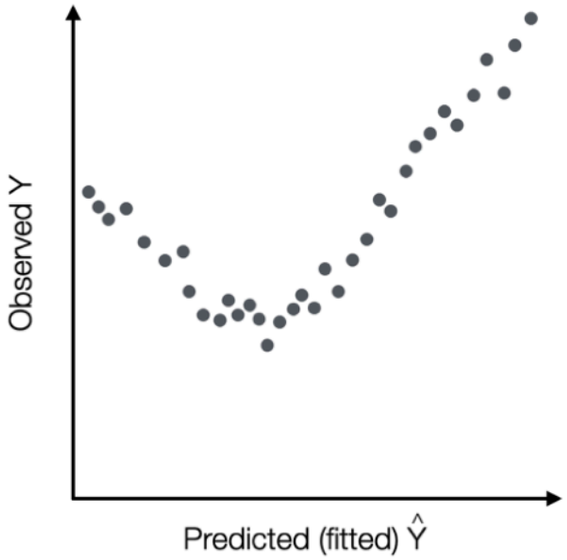
```
> library(car)
> myModel <- lm(outcomeVar ~ varA + varB, data=d)
> Anova(myModel,type=3)
```

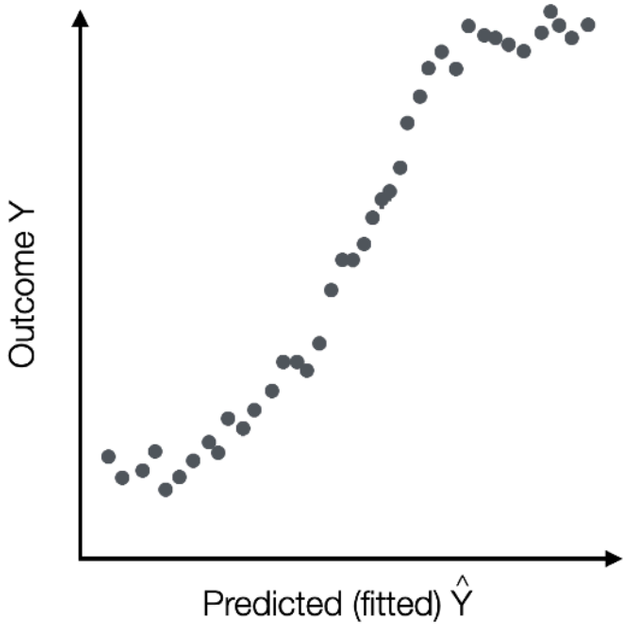
Does a Type III sums of squares.
`type=2` for Type II

See Chapter 16 in *Learning Statistics with R* for lots more details!

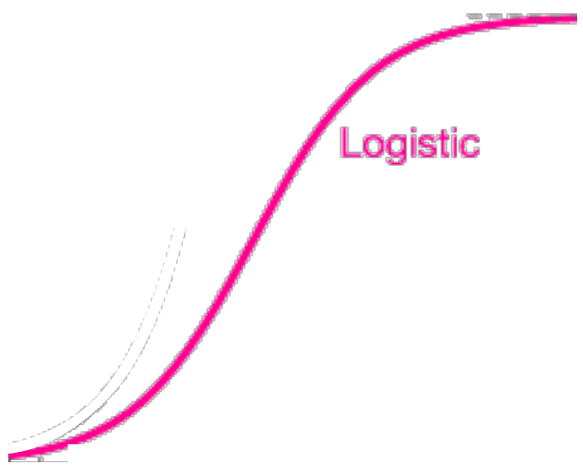
Beyond linear regression

What if you have numeric outcomes and predictors, but the relationship is not linear?



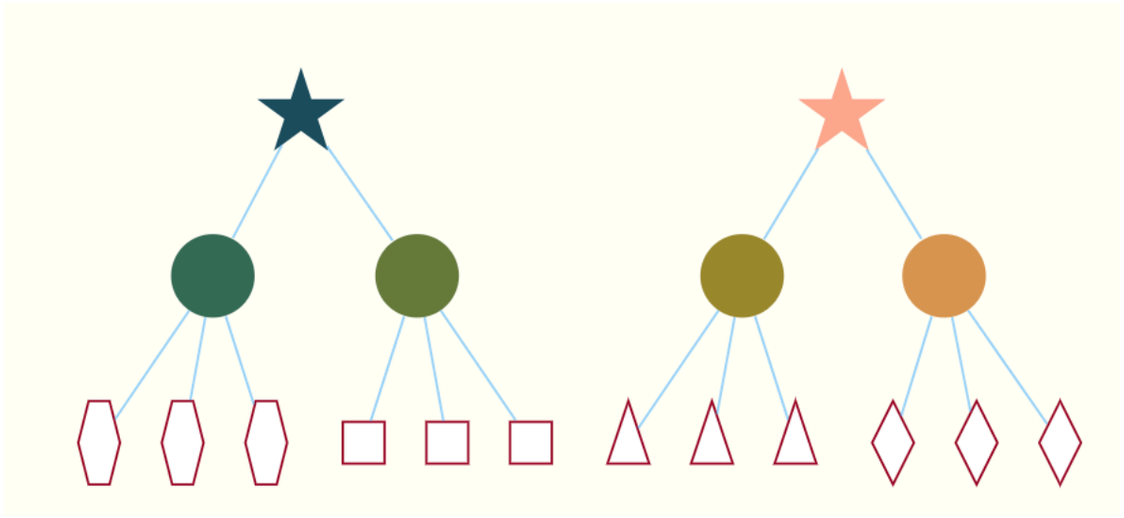


Answer: use other kinds of equation!



Mixed models

Occur when some predictors are nested in others (e.g. each participant contributes multiple trials; you're studying students, each of which occurs in some classroom)



Very useful for making the most out of non-independent data!

The LMS contains some useful resources on how to do these

Bayesian data analysis

Instead of testing a null hypothesis, you compare hypotheses directly by calculating the **likelihood** of the data under each and combining it with its **prior** probability

$$\frac{P(H_2|D)}{P(H_1|D)} = \frac{P(D|H_2)\,P(H_2)}{P(D|H_1)\,P(H_1)}$$

Posterior probability that each hypothesis is true, given the data

Likelihood that the data is true, given the hypothesis

Prior probability that the hypothesis is true

See Week 9 Seminar, Chapter 17 of *Learning Statistics with R*, or *Statistical Rethinking* by McElreath

Wow. That’s a lot of stuff

In the last video I will try to give a bit more of a “forest” view of all of this, rather than the trees we’ve been mired in lately