Analysis of variance and regression

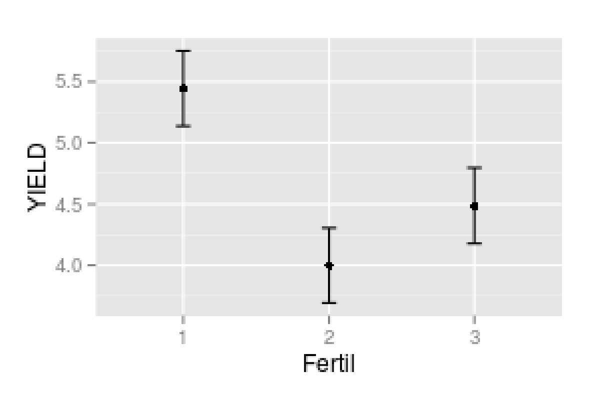
A *review* of the basics

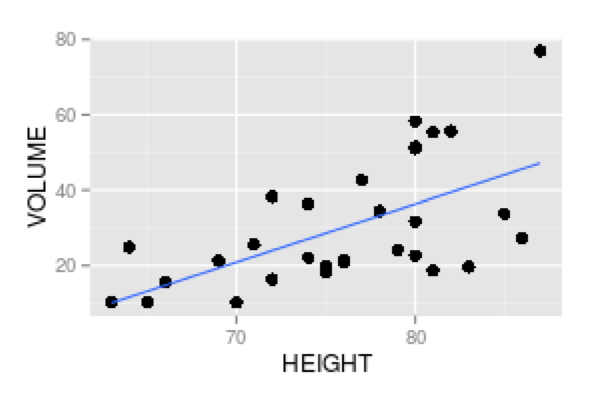
ANOVA and regression

- You learned one-way ANOVA and simple linear regression in your intro stats class
- They were taught as two separate methods
- Today we will review them as you were taught them
- Next week we will start to learn an approach that reveals both to be special cases of a single General Linear Model

What do we want to know about our experimental data?

- When the data are numeric, there are two common questions:
  - Are the means different between groups?
  - Does changing one numeric variable affect another?
- These different questions are addressed with two different statistical analyses

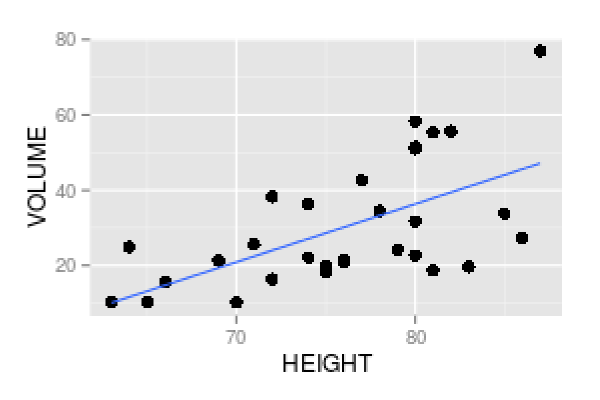




Analysis of effect of one numeric variable on another: linear regression

- Used to measure the straight-line relationshiop between the variables
- Focus is on the properties of the best-fit line (the slope), and the strength of the relationship (coefficient of determination, r^2)

HEIGHT	VOLUME
70	10.3
65	10.3
63	10.2
72	16.4
81	18.8
...	...



Simple linear regression

- Simple linear regression estimates the straight-line relationship between two continuous variables
- One variable is the **independent** (or predictor, x)
  - Experimentally set, or measured without error, or the cause of change
- The other variable is the **dependent** (or response, y)
  - The measure of response to changes in the predictor variable
- Straight line equation: $y = mx + b$
- Regression equation: $\hat{y} = \beta x + \alpha$

Examples of regression questions

- Do more brightly colored birds have more parasites?
- How much lumber is there in a live tree of a particular height?
- How is pest infestation late in the season affected by the concentration of insecticide applied early in the season?

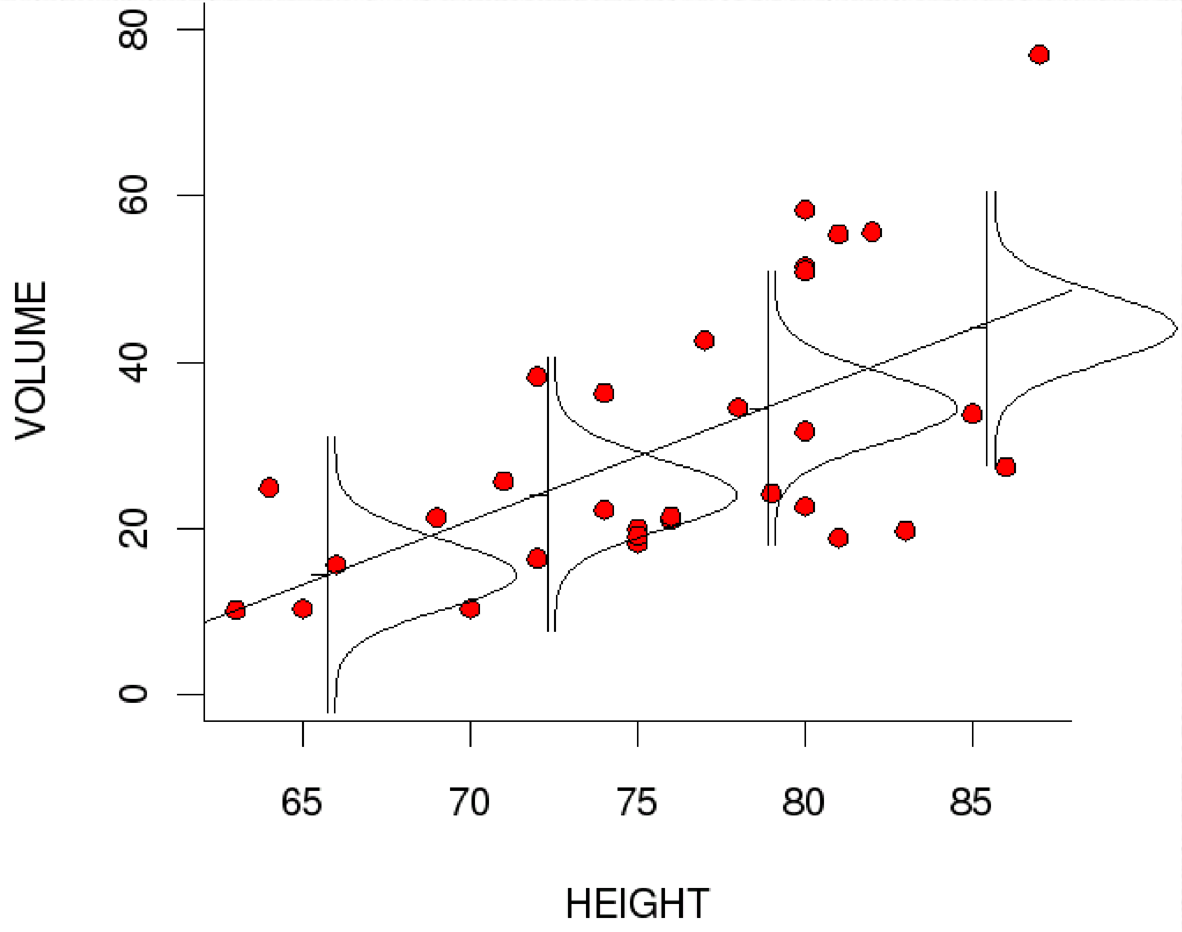
What is the independent (predictor) variable for each?

What the line predicts

Relationship of lumber volume to height of live tree

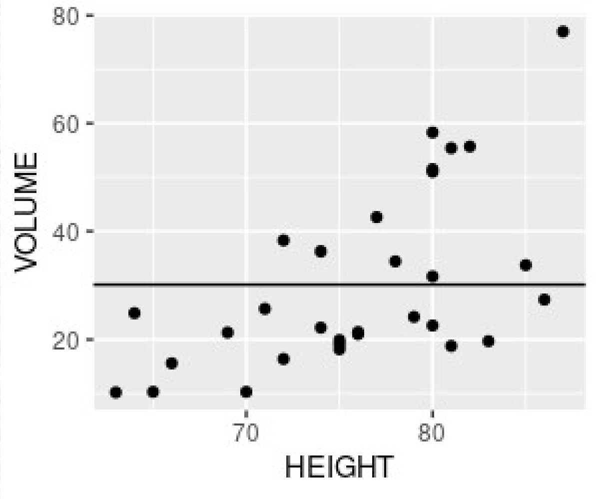
The regression line predicts the ***expected value*** *= predicted mean volume for a given tree height*

Symbolized with a hat, $\hat{y}$ *, to indicate that it's a predicted value*



Testing the statistical significance of a regression line – step 1: the null hypothesis

- Null hypothesis is that there is no relationship between y and x = they are **independent**
- If this is true, knowing x gives you no information about the mean of y
- For any value of x the best prediction possible is the mean of y, $\bar{y}$
- Since $\bar{y}$ is a single number for a given data set, the line is flat – slope of 0



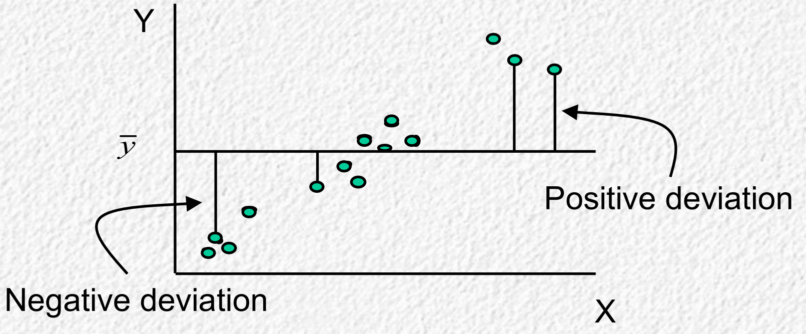
- So, the null is that at the population level the slope, β, is 0 (Ho: β = 0)

Step 2: calculate a test statistic

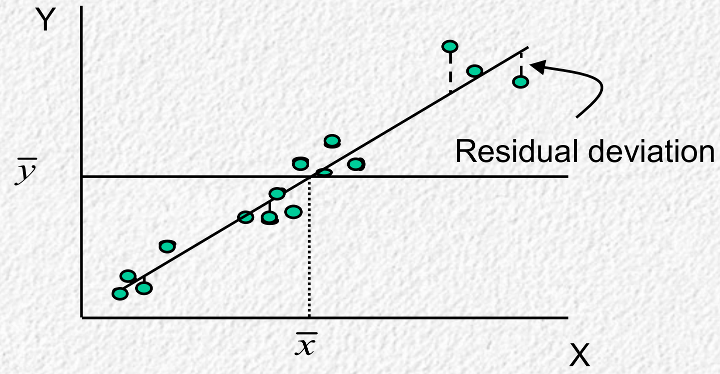
- To the extent that the predictor is causing a change in the response, the best-fit line will predict the data well
  - The line will be in the middle (it's the mean)
  - Data points will vary at random around it
- We want to know if the amount of variation in the data explained by the line is big compared to random variation
- We need to estimate a) how much variation the line explains, and b) how much random variation the line does not explain
- We measure variation with variance = $\Sigma(y_i - \hat{y})^2/(n - 1)$ = SS/df

Partitioning SS in a regression

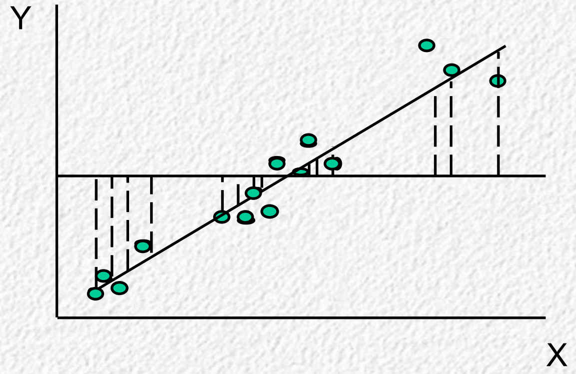
Total SS = sum of squared deviations from observations to mean of y (SST, same as ANOVA)



Residual SS = unexplained variation, sum of squared residuals, SSE (like error SS)

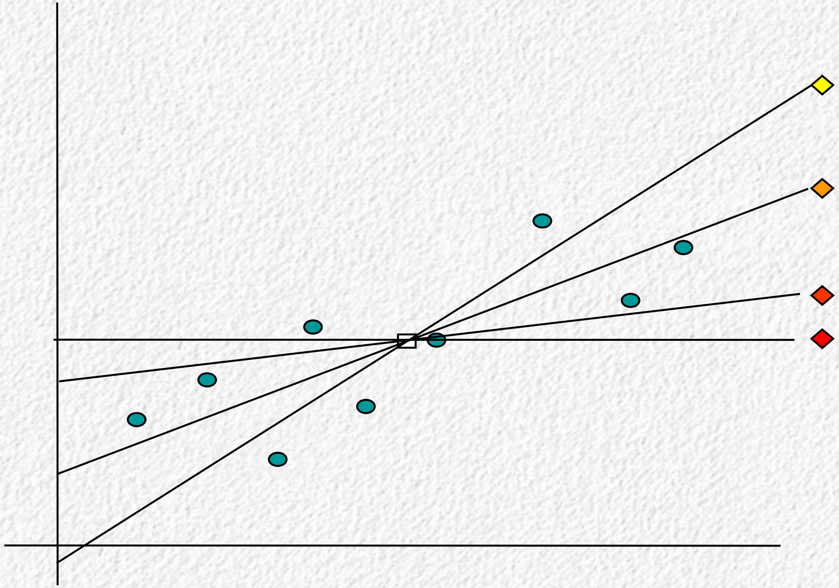


Regression SS = amount of variability attributable to the straight-line relationship, SSR (like SSF).



Because SST = SSR + SSE,
SSR = SST- SSE

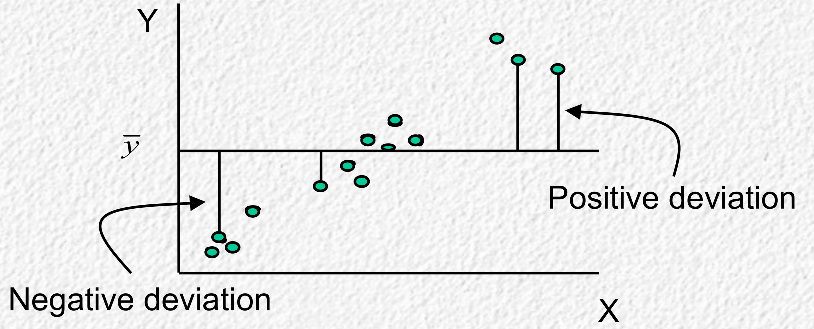
What is the best line?



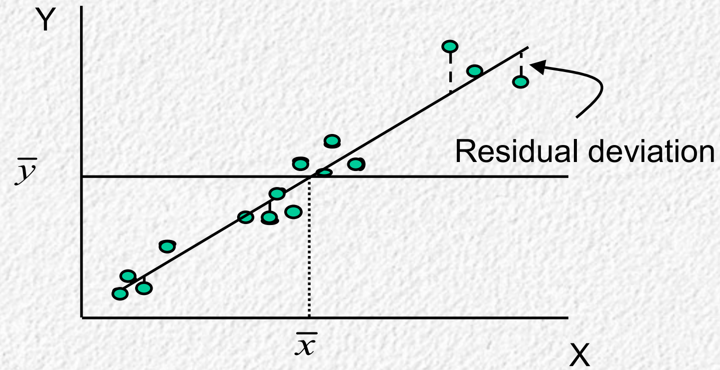
- We want the best line, but many are possible
- The best fit line is considered to be the one that:
  - minimizes the residual SS (the **least squares** criterion)
  - is most likely to have produced the observed data (the **maximum likelihood** criterion)
- When the data are normally distributed, the least squares and maximum likelihood solutions are the same

Degrees of freedom

Total DF = number of measurements – 1 (for estimate of slope)

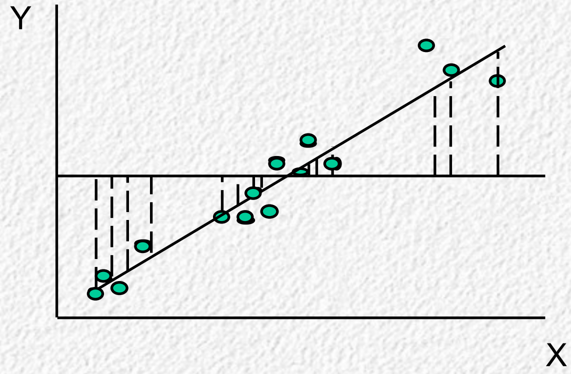


Residual DF = total DF - 1



Regression DF = 1 (for the slope)

Total DF = residual DF + regression DF



Constructing the test statistic: F ratio

- If we divide regression SS by regression DF we get an estimate of variance explained by the line
  - Called the regression **mean squares** ($MS_{regression}$)
- If we divide the residual SS by residual DF we get an estimate of random variance that is not explained by the line
  - Called the residual mean squares ($Ms_{residual}$)
- If the null is true the best fit line is flat, and there is nothing but random variation – $MS_{regression}$ doesn’t account for anything, and will be 0
- So, if we divide $MS_{regression}$ by $MS_{residual}$, bigger values of this ratio give us evidence that we are explaining variation in the response with the predictor
- This ratio, Msregression/Msresidual, is the F ratio – our test statistic

Regression output

MINITAB OUTPUT FOR BOX 2.1 **Analysis of the *trees* dataset: regression**

Regression equation

Test of the coefficients

Regression Analysis: VOLUME **versus** HEIGHT

The regression equation is
VOLUME = - 87.1 + 1.54 HEIGHT

Predictor	Coef	SE Coef	T	P
Constant	-87.12	29.27	-2.98	0.006
HEIGHT	1.5433	0.3839	4.02	0.000

S = 13.40   R-Sq = 35.8%   R-Sq (adj) = 33.6%

Test of the model

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2901.2	2901.2	16.16	0.000
Residual Error	29	5204.9	179.5		
Total	30	8106.1			

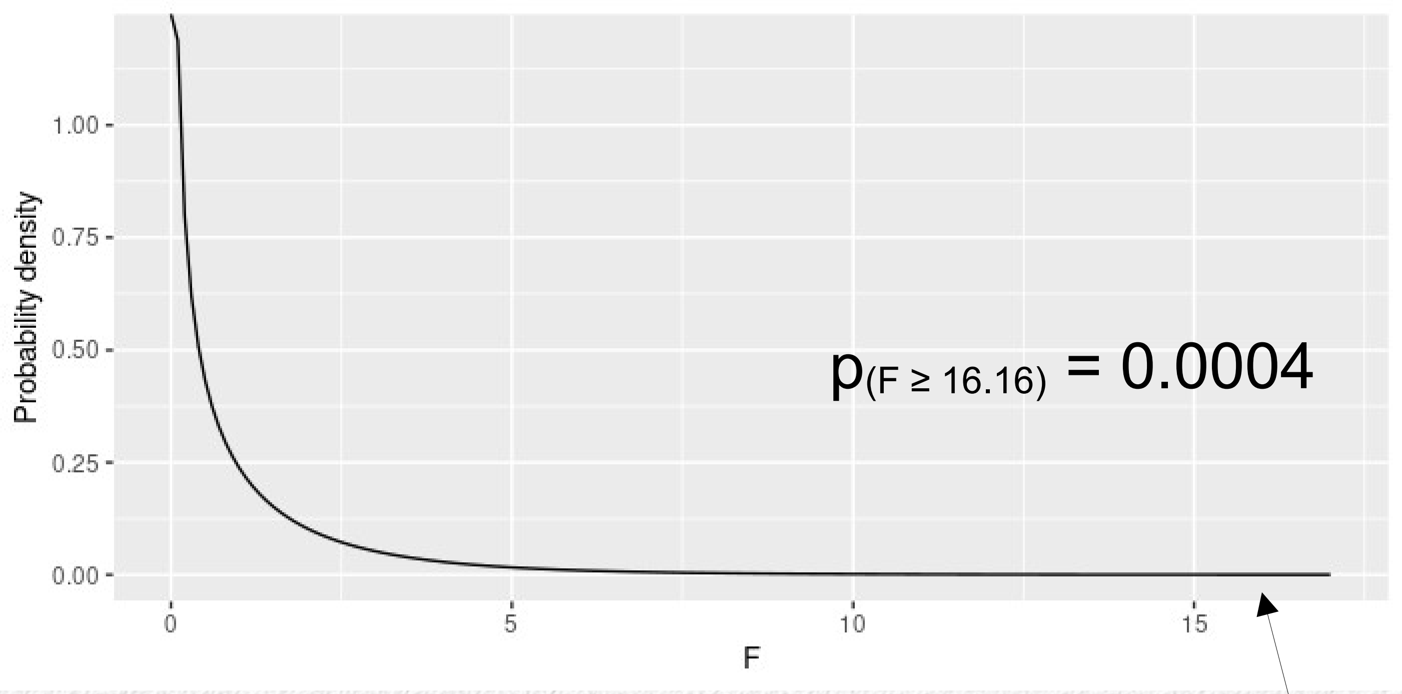
Residual / $\sqrt{MSE}$

Unusual Observations

Obs	HEIGHT	VOLUME	Fit	SE Fit	Residual	St Resid
31	87.0	77.00	47.15	4.86	29.85	2.39R

R denotes an observation with a large standardized residual

P-value: compare F to F distribution



F = 16.16

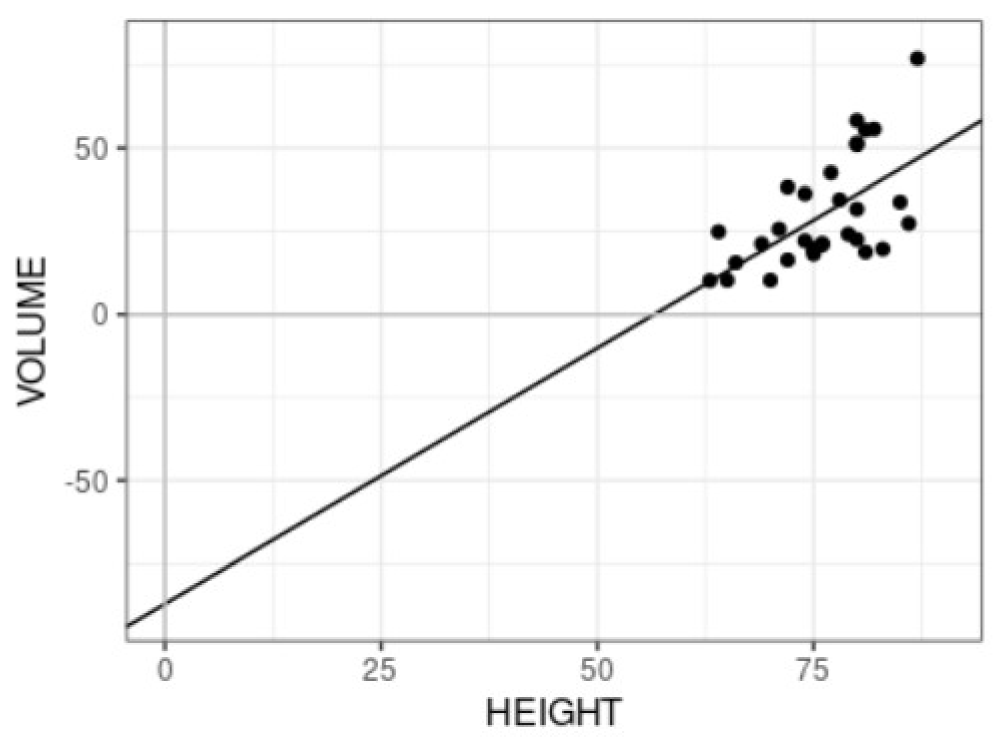
Interpreting the regression

- Interpretation = deriving meaning from the results
- There are two main things we use to interpret regression:
  - The slope of the line
  - How strong the relationship between predictor and response is

Regression produces both slope and intercept coefficients, but we focus on slope

- The regression equation is:

  VOLUME = -87.1 + 1.54 HEIGHT
- The intercept is the volume expected when height = 0
  - Outside of measured range of data
  - Ridiculous estimate
  - Doesn't measure cause and effect relationship
- Slope is ΔVOLUME/ΔHEIGHT (rise/run)
  - Direct measure of how the response (volume) changes with each 1 unit change of the predictor (height)
  - Measures the cause/effect relationship, which is what we want to know
- Slope is always interpreted, intercept usually is not in a linear regression



“Significant” is not synonymous with “strong relationship”

MINITAB OUTPUT FOR BOX 2.1 **Analysis of the *trees* dataset: regression**

Regression Analysis: VOLUME versus HEIGHT

The regression equation is
VOLUME = - 87.1 + 1.54 HEIGHT

Predictor	Coef	SE Coef	T	P
Constant	-87.12	29.27	-2.98	0.006
HEIGHT	1.5433	0.3839	4.02	0.000

S = 13.40 R-Sq = 35.8% R-Sq (adj) = 33.6%

Analysis of Variance

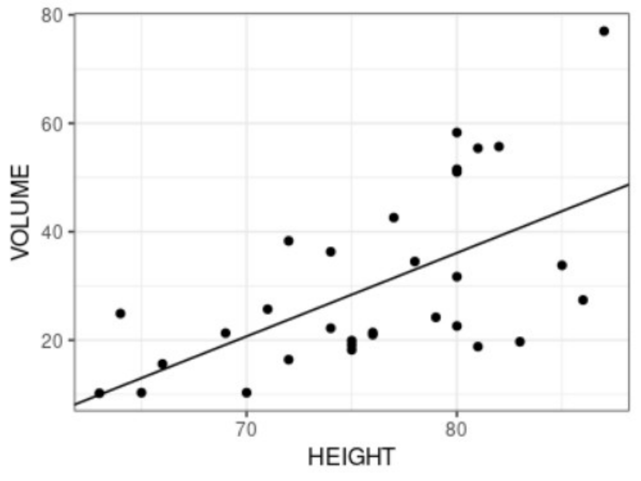
Source	DF	SS	MS	F	P
Regression	1	2901.2	2901.2	16.16	0.000
Residual Error	29	5204.9	179.5		
Total	30	8106.1			

Unusual Observations

Obs	HEIGHT	VOLUME	Fit	SE Fit	Residual	St Resid
31	87.0	77.00	47.15	4.86	29.85	2.39R

R denotes an observation with a large standardized residual

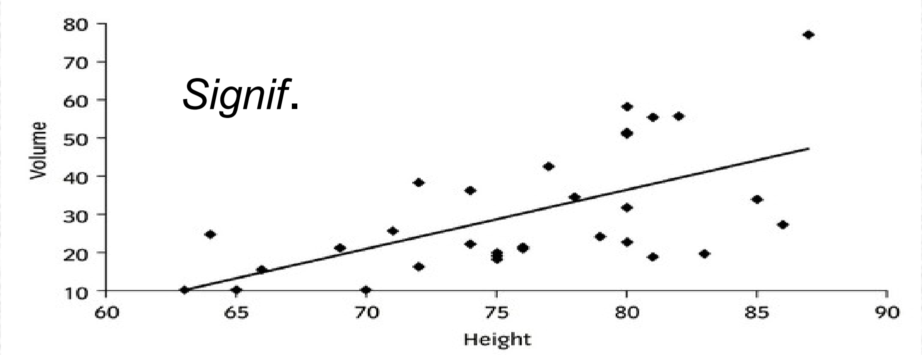
How strong is the relationship? Look at the scatter around the line

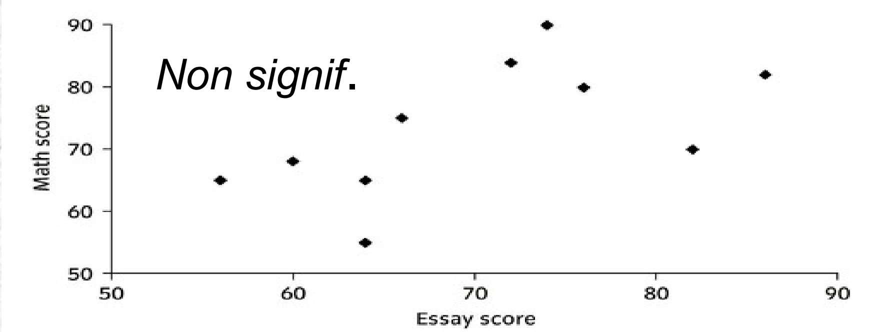


p-values don't reliably indicate strength of relationship

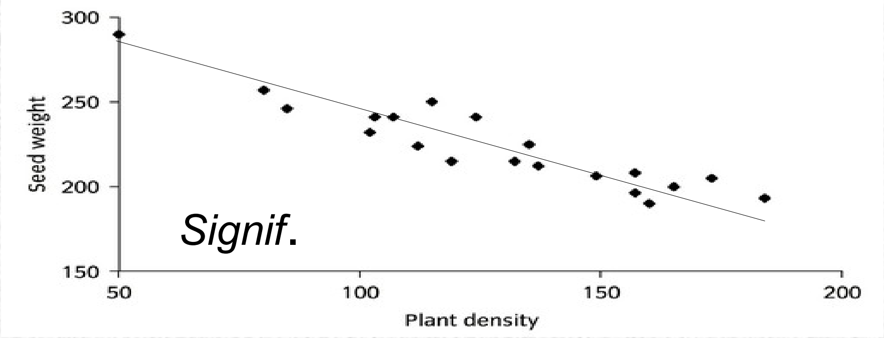
Large sample size | Small sample size

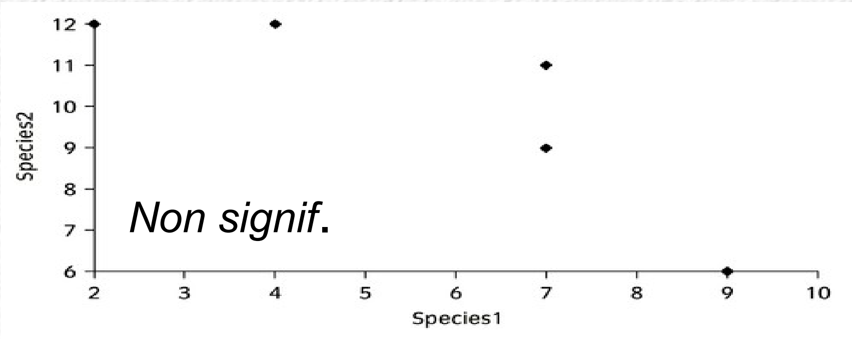
Low r^2



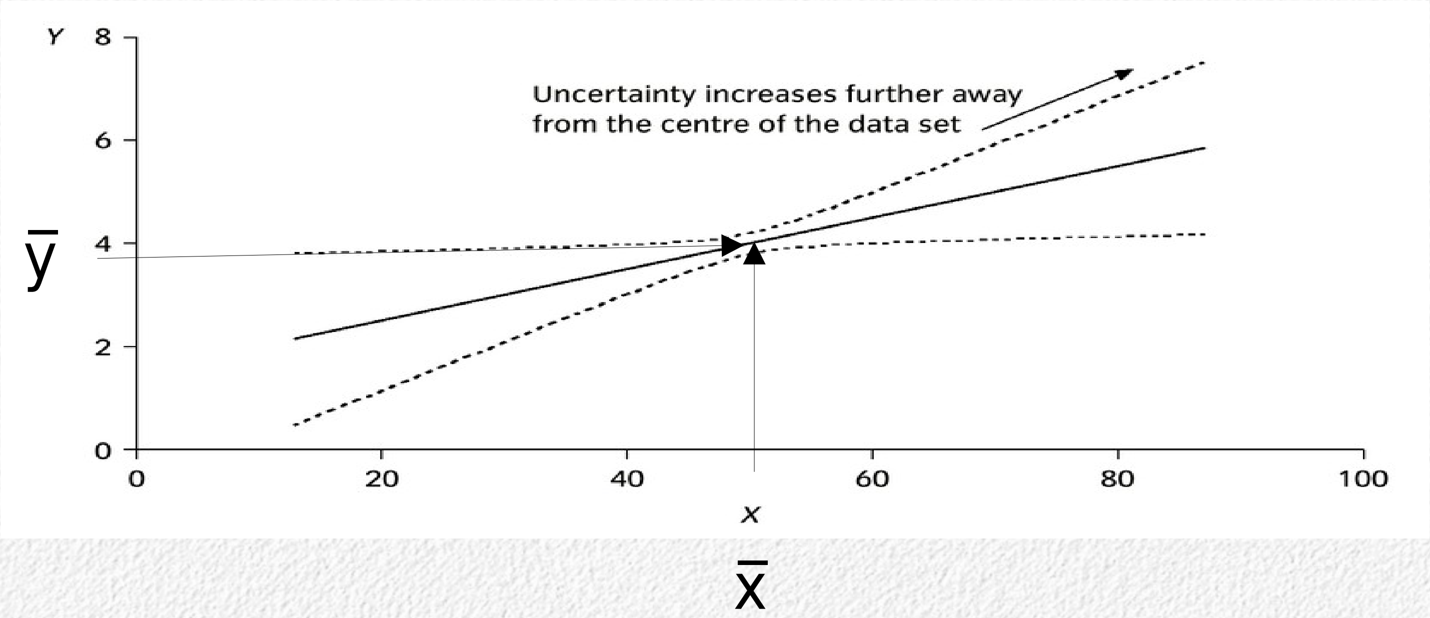


High r^2



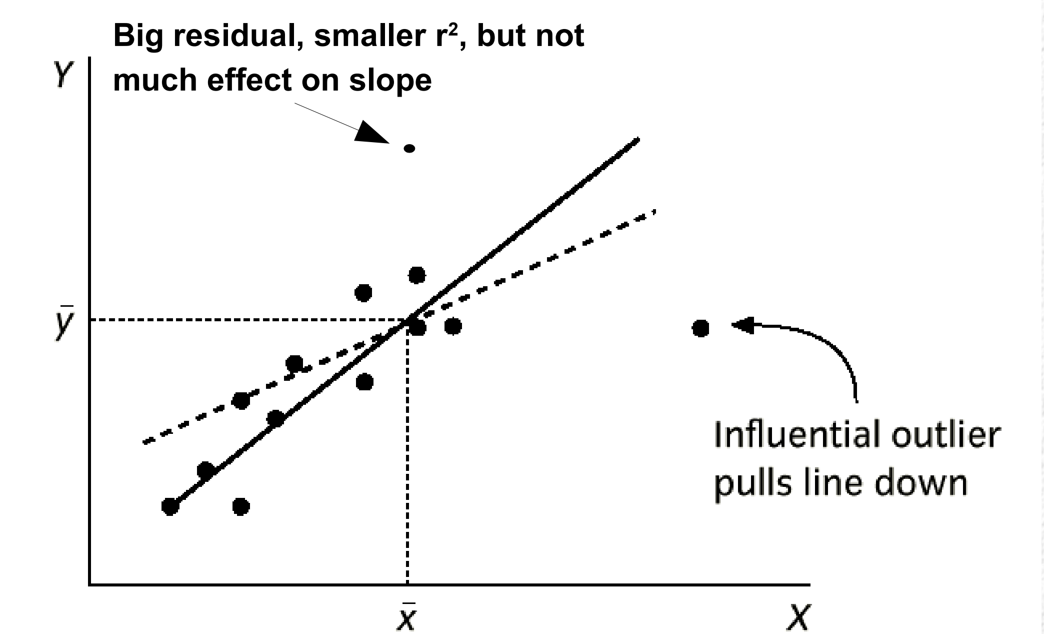


Confidence limits for a regression line



Unusual observations, a.k.a "outliers"

- Outliers can be real, but are often errors
- Examine outliers, fix errors, but beware of throwing them away
- Outliers on Y but not X affect r^2
- Outliers on both Y and X affect both r^2 and the estimate of the slope of the line



There's an app for that...

Comparison of means of groups: ANOVA

- Used for comparisons of 2 or more means (better than t-tests when there are 3 or more means to compare)
- Focus is on amount of difference between means, whether it is large compared with random variation

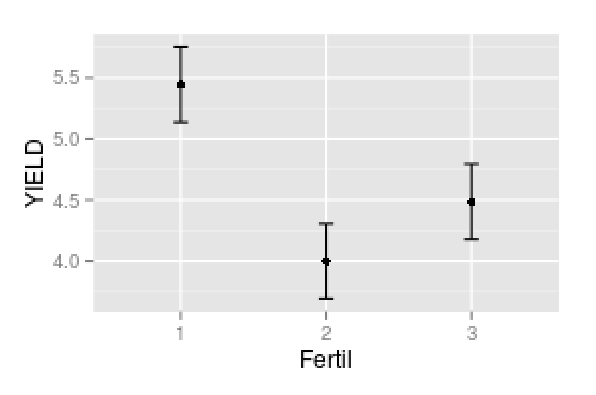


Table 1.1 Raw data from the *fertilisers* dataset

Fertiliser	Yields (in tonnes) from the 10 plots allocated to that fertiliser
1	6.27, 5.36, 6.39, 4.85, 5.99, 7.14, 5.08, 4.07, 4.35, 4.95
2	3.07, 3.29, 4.04, 4.19, 3.41, 3.75, 4.87, 3.94, 6.28, 3.15
3	4.04, 3.79, 4.56, 4.55, 4.53, 3.53, 3.71, 7.00, 4.61, 4.55

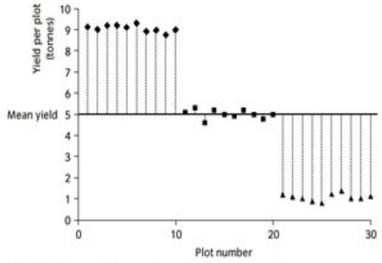
Variances to compare means: partition variation

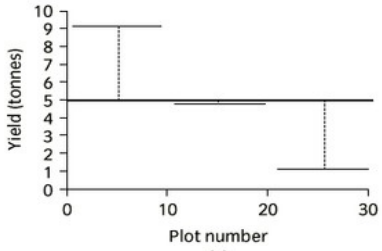
- We don't have a line, but we do have group means
  - Variation between the group means is a measure of the effect of the treatments – if they have no effect, the group means are only different due to random chance
  - Variation around group means is due to individual, random differences between subjects
- So, the equivalent of a regression SS will be obtained via differences between group means and the grand mean (mean of all the y data)
- And the equivalent of residual SS will be obtained from differences between data values and group means

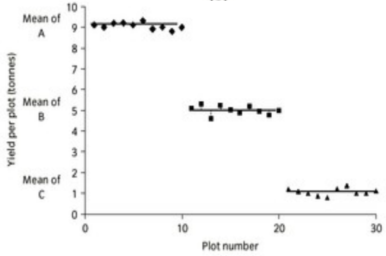
Another app...

Calculations for ANOVA

	SS	d.f.
Total sums of squares (SSY)	$\sum_{i=1}^{an}\left(y_{i,j}-\bar{\bar{y}}\right)^2$	an-1
Fertilizer sums of squares (SSF)	$n\sum_{j=1}^{a}\left(\bar{y}_j-\bar{\bar{y}}\right)^2$	a-1
Error sums of squares (SSE)	$\sum_{j=1}^{a}\sum_{i=1}^{n}\left(y_{i,j}-\bar{y}_j\right)^2$	$df_Y - df_F$







SSY = SSF + SSE

$df_Y = df_F + df_E$

a = number of groups

n = number of observations in each group (groups have same n here)

an = total observations in the data set

Shape of F-distribution depends on both DF

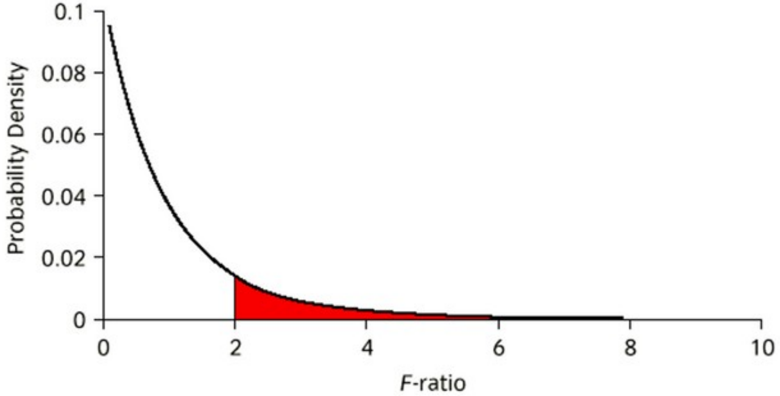
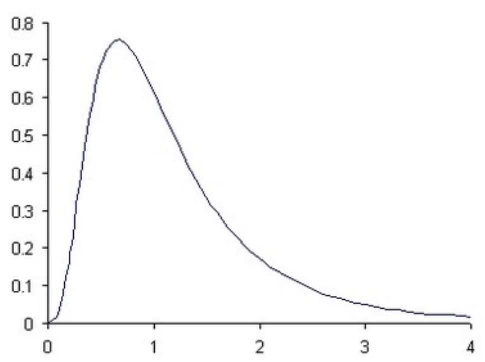


Fig. 1.7 The *F* distribution for 2 and 27 degrees of freedom (illustrates the probability of a *F*-ratio of different sizes when there are no treatment differences).



F with num. df = 10, denom df = 10

F with 2 numerator DF and 27 denominator DF

How many groups? How many data values?

The ANOVA table

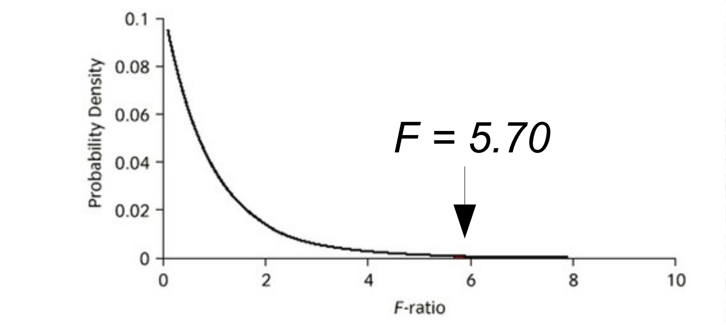


Fig. 1.7 The F distribution for 2 and 27 degrees of freedom (illustrates the probability of a F-ratio of different sizes when there are no treatment differences).

BOX 1.1 Analysis of variance with one explanatory variable

Word equation: `YIELD = FERTIL`

`FERTIL` is categorical

One-way analysis of variance for `YIELD`

Source	DF	SS	MS	F	P
FERTIL	2	10.8227	5.4114	5.70	0.009
Error	27	25.6221	0.9490		
Total	29	36.4449			

EMS = SSE / 27

FMS = SSF / 2

F = 5.41/0.95 = 5.70

p = 0.009

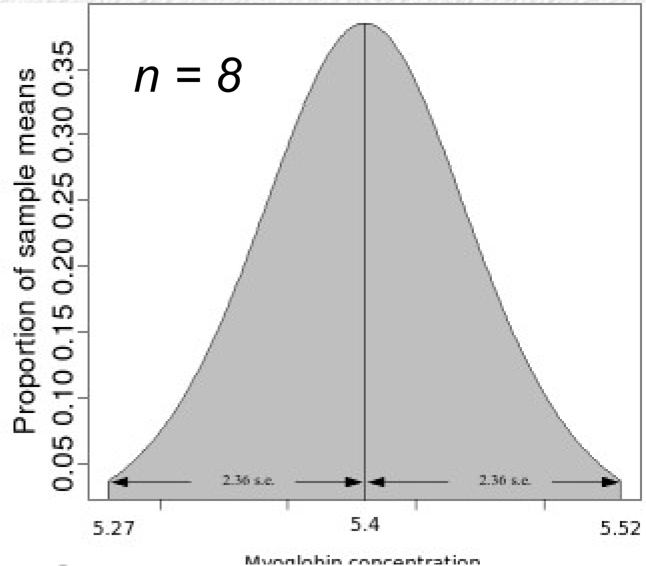
Reject the null, conclude at least one pair of means are different

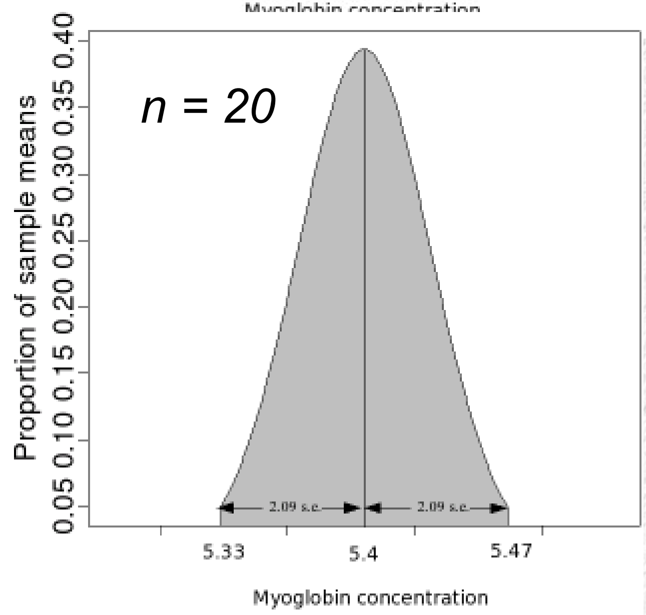
*This is the **omnibus test** of the model – still need to determine which means are different*

Quick review of confidence intervals...

- Sample means are estimates, true (population) value not known
- Confidence intervals put reasonable bounds on what the true value might be, at some level of confidence (usu. 95%), given sampling variation
- Confidence intervals are t standard errors on either side of means
- Standard error is determined by variability of observations and n
- t is determined by d.f. (n) and confidence level

$$Limits = \bar{x} \pm t\, s_{\bar{x}}$$





Standard errors and confidence intervals for group means from an ANOVA

Table 1.4 Constructing confidence intervals

Fertiliser	$\bar{y}$	t_{crit} with 27 df for 95% confidence	$\frac{s}{\sqrt{n}}$	Confidence interval
1	5.445	2.0518	0.3081	(4.81, 6.08)
2	3.999	2.0518	0.3081	(3.37, 4.63)
3	4.487	2.0518	0.3081	(3.85, 5.12)

$$\bar{y} \pm t_{crit} \frac{s}{\sqrt{n}}$$

$$\sqrt{\frac{0.949}{10}}$$

Descriptive Statistics

Variable	FERTIL	N	StDev
YIELD	1	10	0.976
	2	10	0.972
	3	10	0.975

Analysis of Variance for YIELD

Source	DF	SS	MS
FERTIL	2	10.823	5.411
Error	27	25.622	0.949
Total	29	36.445	

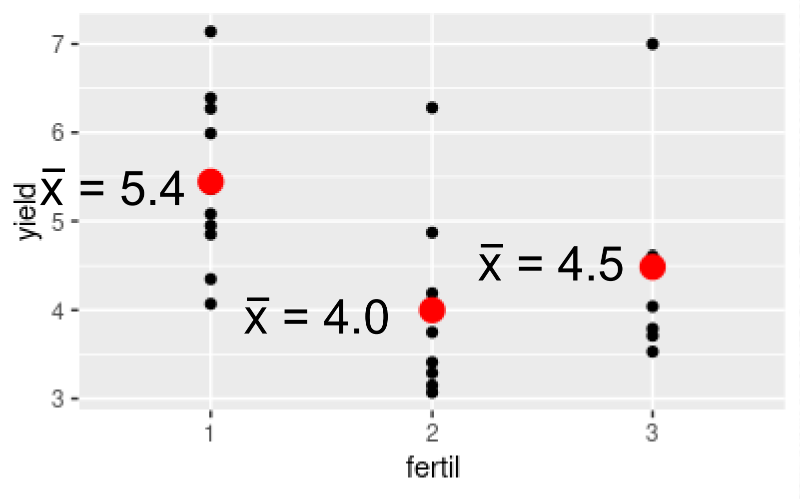
t-distribution: converges on normal, so t-values will usually be about 2 when df is 20 or more

Tukey post-hoc comparisons for fertilizer data

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ fertil, data = fert.dat)

$fertil
      diff        lwr        upr     p adj
2-1 -1.446 -2.5261662 -0.3658338 0.0070788
3-1 -0.958 -2.0381662  0.1221662 0.0894812
3-2  0.488 -0.5921662  1.5681662 0.5102335
```



Summary: ANOVA and regression

- Each handles a different type of data:
  - Grouped data = ANOVA
  - Two numeric variables: regression
- Both use a numeric response variable
- Regression uses a numeric predictor, and ANOVA uses a categorical predictor (Fertilizer, levels = 1, 2, 3)
- Both test statistical significance by partitioning variance
- Not a coincidence! Both are special cases of the General Linear Model, which we will learn about nest week!