Multiple regression

Regression with more than one predictor

Multiple regression – using 2 or more predictors for a response

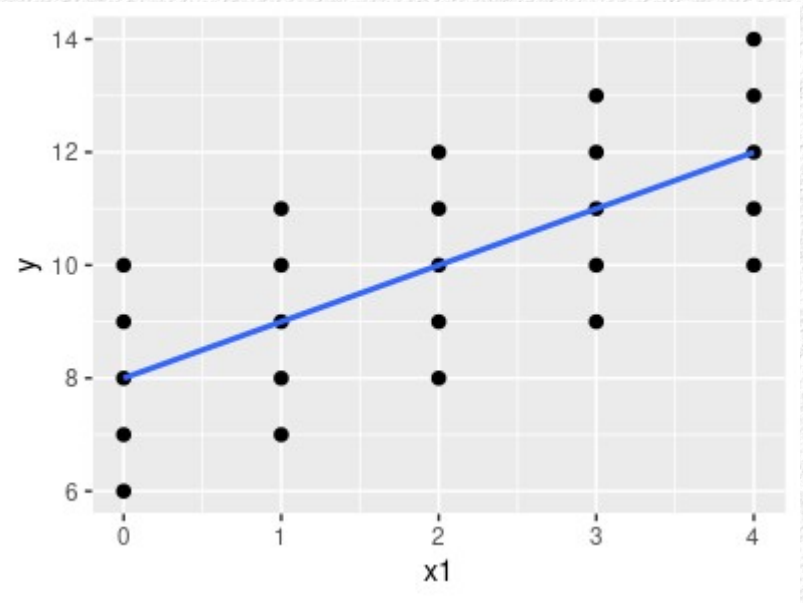
- Linear combination
  - Linear = a coefficient multiplied by a variable
  - Combination = terms added together

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

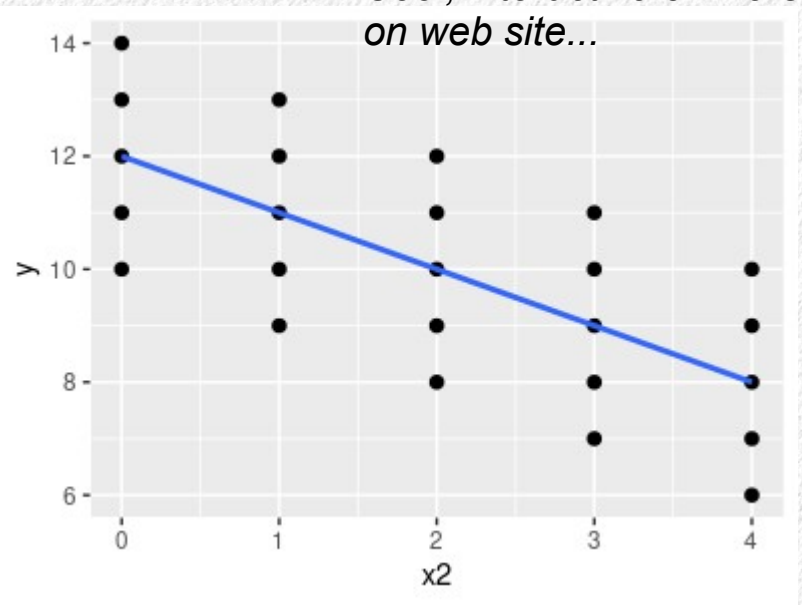
- Model predicts position on a plane, instead of a line (if 3 predictors it's a hyperplane)
  - Or, you can think of it as a series of parallel lines
- Additional predictors added as coefficients multiplied by the predictor variable

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n$$

Two predictors, one at a time



Cool, interactive 3D version on web site...



$$\hat{y} = 8 + 1\,x_1$$

$$\hat{y} = 12 - 1\,x_2$$

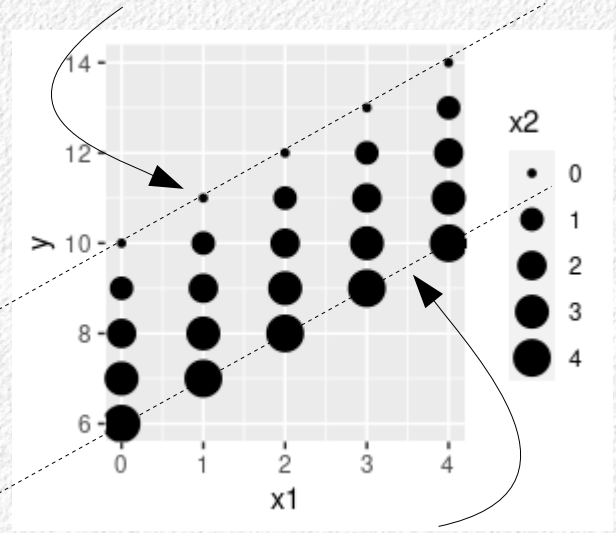
r^2 = 0.5 for each, one at a time

Holding x_2 constant, varying x_1

$$\hat{y} = 10 + 1\,x_1 - 1\,x_2$$

x2	x1	y
0	0	10
	1	11
	2	12
	3	13
	4	14
4	0	6
	1	7
	2	8
	3	9
	4	10

$$\hat{y} = 10 + 1\,x_1 - 1(0) = 10 + 1\,x_1$$



$$\hat{y} = 10 + 1\,x_1 - 1(4) = 6 + 1\,x_1$$

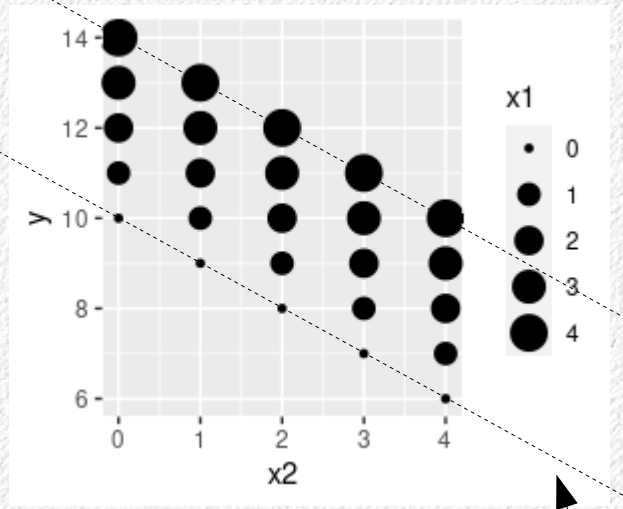
And, if we continued by setting x2 to 1, 2, and 3 we predict every point exactly – $R^2 = 1$ together

Holding x_1 constant, varying x_2

$$\hat{y} = 10 + 1\,x_1 - 1\,x_2$$

x1	x2	y
0	0	10
	1	9
	2	8
	3	7
	4	6
4	0	14
	1	13
	2	12
	3	11
	4	10

$$\hat{y} = 10 + 1(4) - 1\,x_2 = 14 - 1\,x_2$$



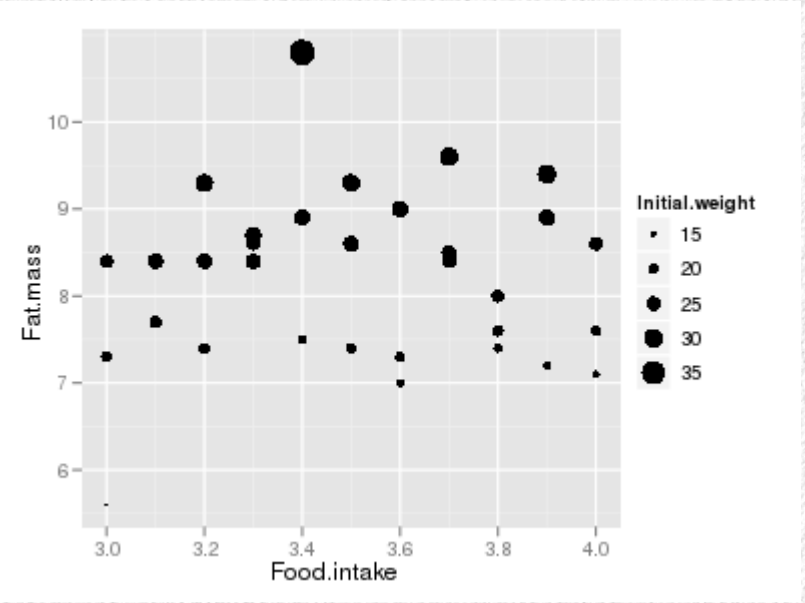
$$\hat{y} = 10 + 1(0) - 1\,x_2 = 10 - 1\,x_2$$

Reasons to use more than one predictor

- Statistical elimination – accounting for one (nuisance) variable so that the effect of another can be measured
  - Reducing statistical noise – a nuisance variable affects the response and needs to be accounted for
    - Example: variation in initial mass masking the effect of food intake on fat mass of mice
  - Avoiding spurious relationships – a relationship that appears to be due to one predictor is actually due to another
    - Example: apparent effect of height on math test scores in kids
- Accounting for complexity – more than one predictor necessary to understand the response
  - Example: wood volume as a function of tree height and diameter

Fat mass by food intake

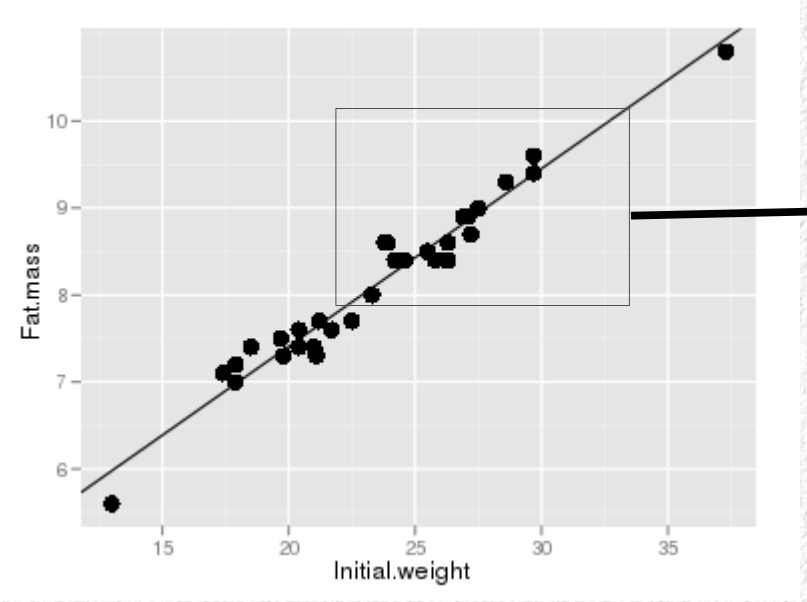
- Fat mass of mice given different daily food intakes
- Noisy data!
  - Mice didn't all start at the same size
  - Leads to lots of variation in fat mass that isn't due to food intake
- Not significant (p = 0.65)
- How can we reduce the noise?



We can’t fix this problem in our design

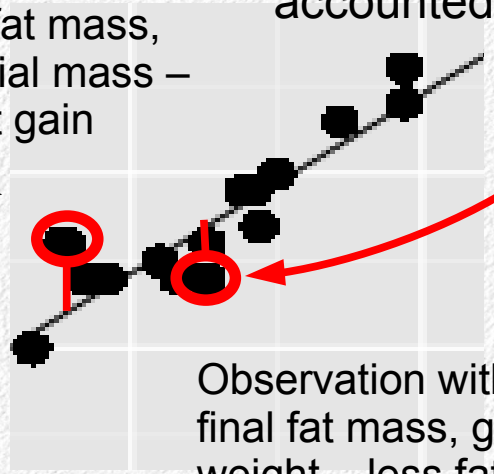
- Ideal experiment would:
  - Use animals that were all the same size, with the same initial fat masses
  - Measure the initial fat mass and the final fat mass, use the difference to measure fat gain
- Real experiments not always ideal
  - Mice come in different sizes, with different amounts of fat – can’t just set them all to the same initial sizes
  - Measuring fat mass accurately is done by extracting fat from a carcass, can only be done once
- Fixing the problem by changing the experimental design is not possible
- We need to address the problem analytically

Relationship between initial weight and fat mass



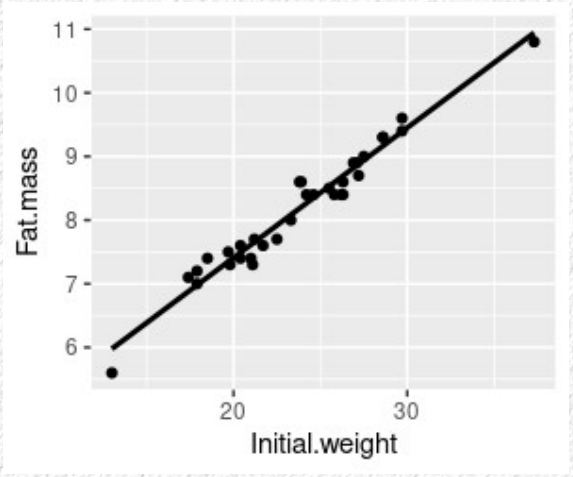
Observation with a big final fat mass, given initial mass – lots of fat gain

Residual is variation left after initial weight is accounted for

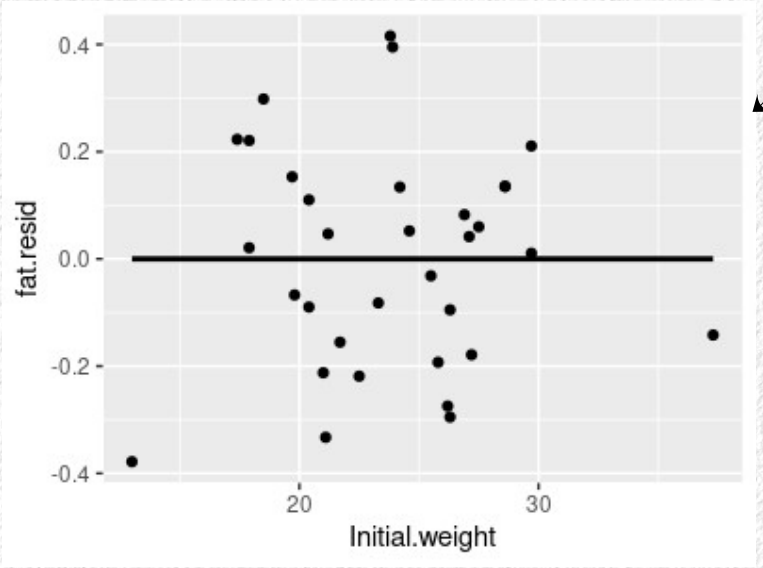


Observation with a small final fat mass, given initial weight – less fat gain (maybe loss)

$$Fat\ .\ mass = 3.32 + 0.20\ Initial\ .\ weight$$



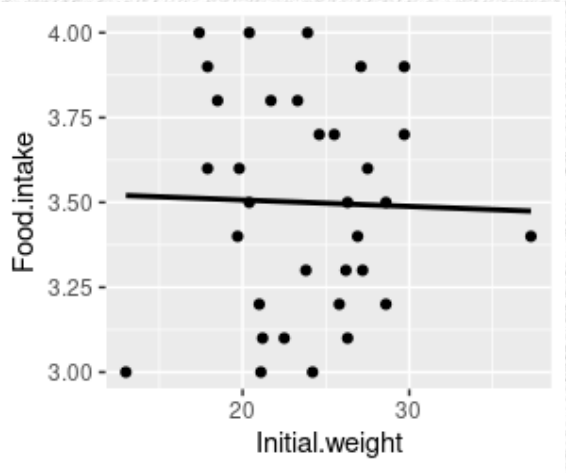
Residuals are the fat mass independent of initial weight

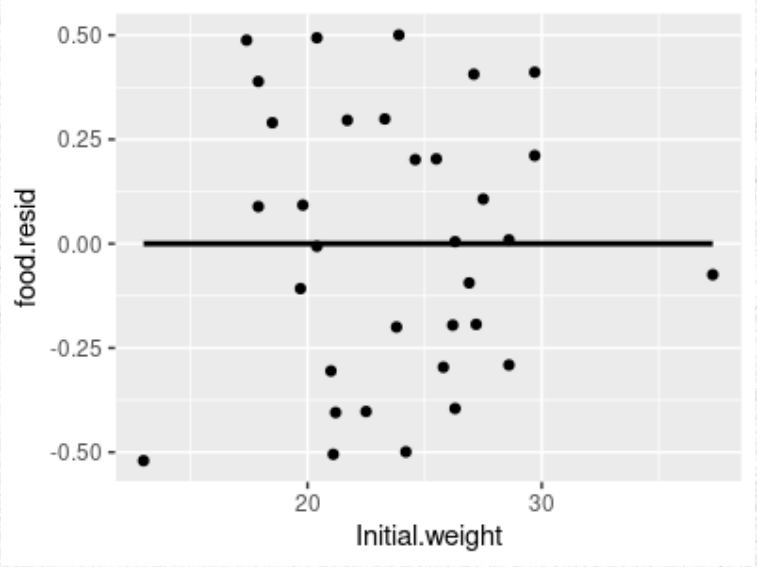


The fat vs. initial weight regression line is the relationship between the variables

Residuals are variation in fat that's unrelated to initial weight

Like setting all the animals to the same starting weight of 0





Also need food intake independent of initial weight

Correlation between food intake and initial weight is very small

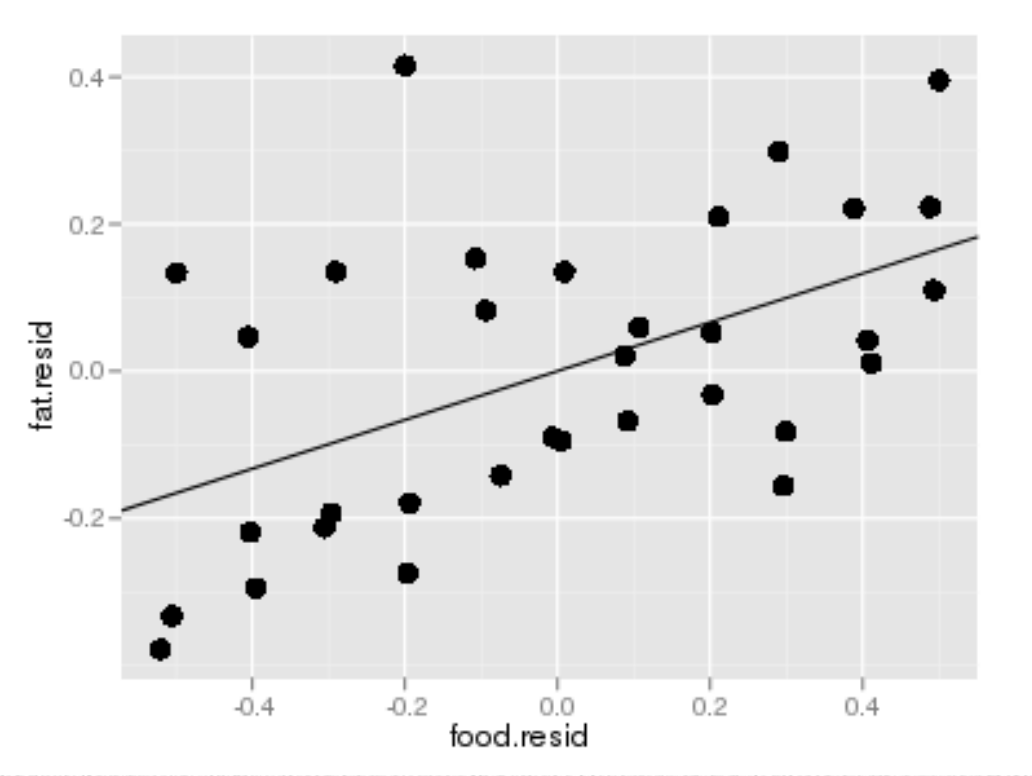
(slope (b) = -0.00189, r = -0.028)

Need the relationship to be 0

Residuals for this relationship make food intake perfectly independent of initial weight
(r = 0, slope = 0)

Relationship between fat and food intake, with initial weight eliminated

- Regress fat residual on food residual
- Now can see the positive relationship
- Significant ($p = 0.001$)
- Slope is 0.331



Another cool 3D graph...

Simpler (better) approach: include initial weight as a second predictor

```
Response: Fat.mass
                Df  Sum Sq Mean Sq F value    Pr(>F)
Initial.weight   1 29.5886 29.5886 926.837 < 2.2e-16
Food.intake      1  0.3628  0.3628  11.363  0.002076
Residuals       30  0.9577  0.0319
```

Both initial weight and food intake are predictors in the model

Each is tested for an effect on fat mass

F is $MS_{predictor}/MS_{residual}$ *for each*

Correctly accounts for the fact that the relationship with initial weight is estimated from the data – 1 df deducted for initial weight

Model multiple R^2

```
Response: Fat.mass
                Df  Sum Sq Mean Sq F value    Pr(>F)
Initial.weight   1 29.5886 29.5886 926.837 < 2.2e-16
Food.intake      1  0.3628  0.3628  11.363  0.002076
Residuals       30  0.9577  0.0319
```

Now a multiple R^2 – tells us how much variation is explained by the entire model

Here calculated as sum of the two predictor SS (works with Type I ANOVA)

$$Model\ R^2 = \frac{(29.5886 + 0.3628)}{(29.5886 + 0.3626 + 0.9577)} = 0.97$$

Coefficients tested are ***partial*** regression coefficients

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.146942   0.384845   5.579 4.55e-06 ***
Initial.weight 0.204894   0.006712  30.526  < 2e-16 ***
Food.intake  0.331684   0.098394   3.371  0.00208 **
```

Slope on food intake identical to fat resid on food resid, but including initial weight as a predictor properly accounts for estimating its slope – deducts 1 df from residuals

Intercept is fat mass expected when initial weight and food intake both set to 0

Cost of failing to account for initial weight

```
Response: Fat.mass
                Df  Sum Sq Mean Sq F value    Pr(>F)
Initial.weight   1 29.5886 29.5886 926.837 < 2.2e-16 ***
Food.intake      1  0.3628  0.3628  11.363  0.002076 **
Residuals       30  0.9577  0.0319
```

```
Response: Fat.mass
                Df  Sum Sq Mean Sq F value    Pr(>F)
Food.intake      1  0.3628  0.3628  0.3682  0.548412
Residuals       31  30.546  0.9854
```

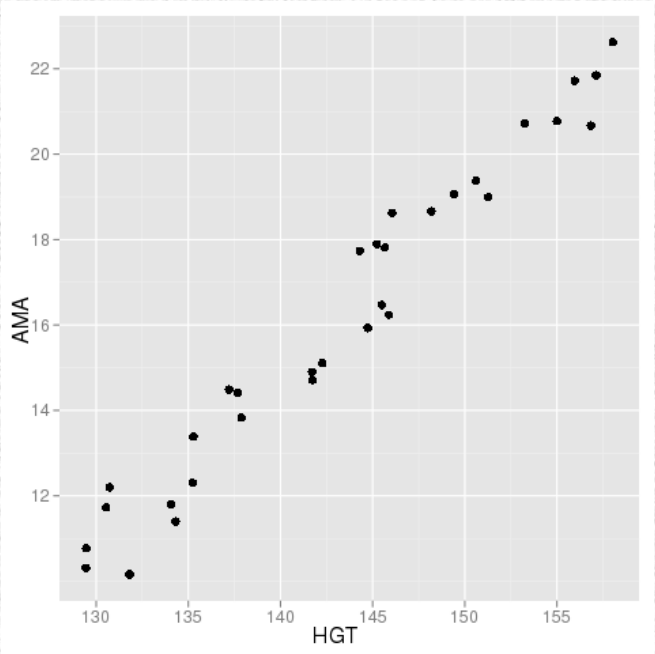
Variation that isn't accounted for goes into the residual term → smaller F, bigger p

Same amount of explained variance for food intake is no longer significant

Spurious relationships

- Relationships that are statistically significant, but don't represent a cause and effect relationship, are “spurious”
- They often happen because a third variable is actually responsible
  - The third variable is responsible for a change in both the predictor and the response → predictor and response appear to be related
- How can we know?

Example: nuisance variables and spurious relationships



AMA	HGT
10.31	129.44
10.77	129.46
10.16	131.81
11.73	130.54
12.20	130.73
11.40	134.31
11.80	134.07
13.39	135.27
12.30	135.24
...	...

AMA – scores by grade school kids on a standardized math test

Why might this be spurious? What might be a better explanation?

Both height and math ability are related to age

Fig. 4.2(a) HGHT against YEARS.

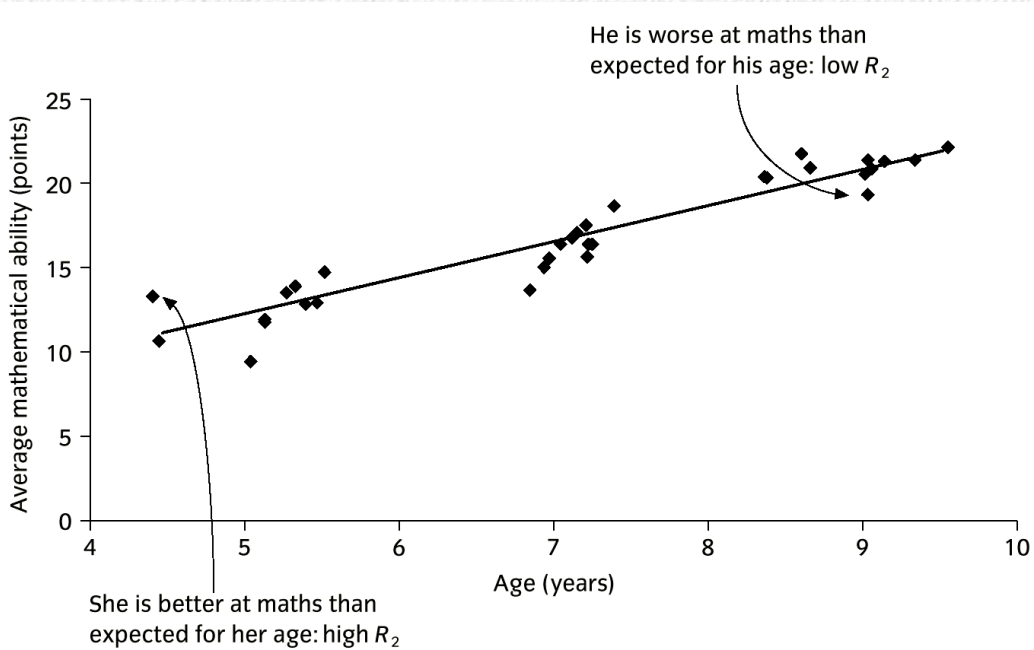
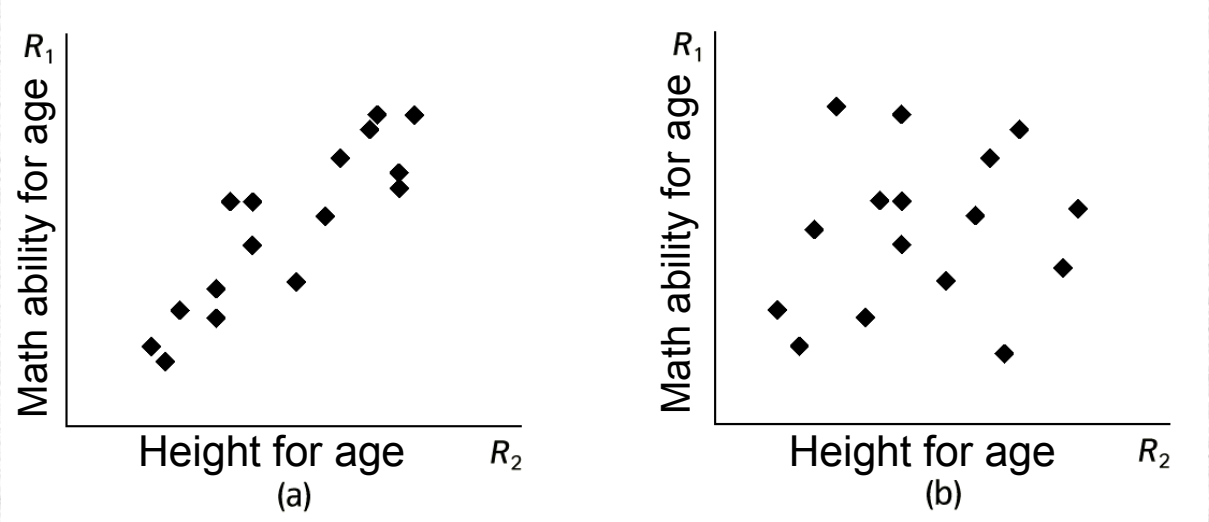


Fig. 4.2(b) AMA against YEARS.

In 3D...

Is there still a relationship between height and AMA, once age is accounted for?



One possibility: still a relationship after accounting for age

Other possibility: all of the apparent relationship was due to age, none left after age is accounted for

Which one is it?

No effect of height once age is accounted for

BOX 4.1 Height explaining mathematical ability

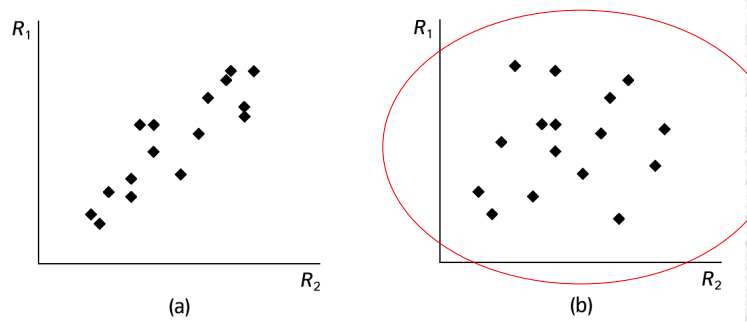
General Linear Model

Word equation: AMA = HGHT

HGHT is continuous

Analysis of variance table for AMA, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
HGHT	1	412.77	412.77	412.77	726.87	0.000
Error	30	17.04	17.04	0.57		
Total	31	429.81				



The relationship between height and math ability is spurious

BOX 4.2 Years, not height, explaining mathematical ability

General Linear Model

Word equation: AMA = YEARS + HGT

YEARS and HGT are continuous

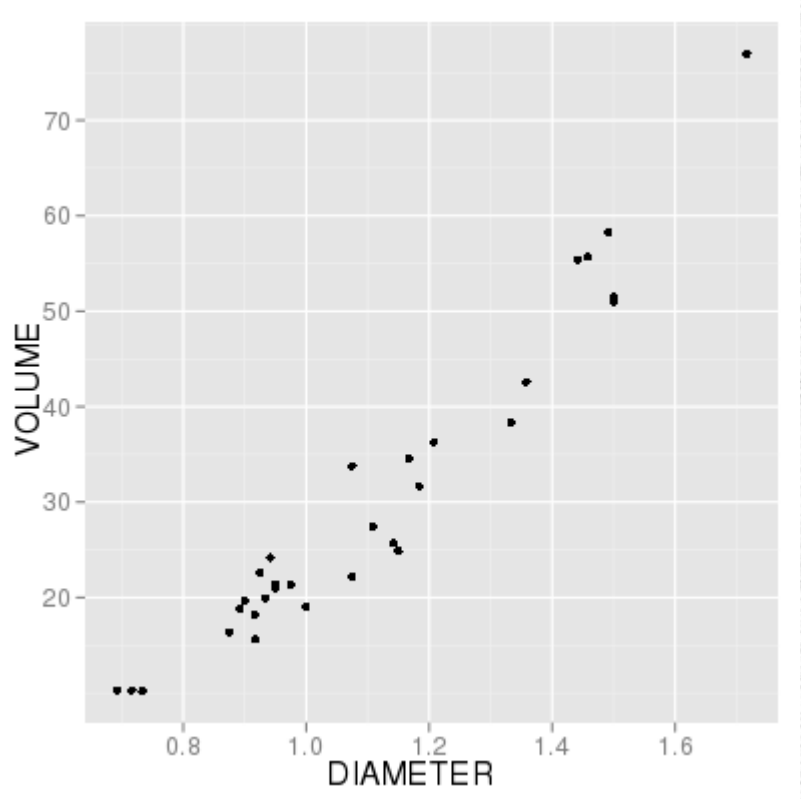
Analysis of variance table for AMA, using Adjusted SS for tests

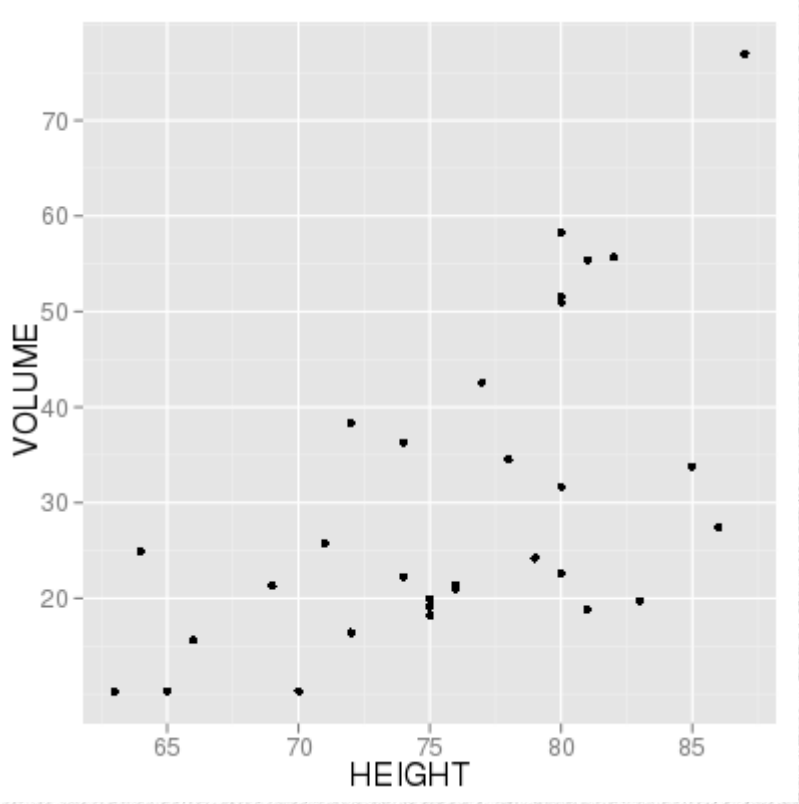
Source	DF	Seq SS	Adj SS	Adj MS	F	P
YEARS	1	422.60	9.84	9.84	39.63	0.000
HGT	1	0.01	0.01	0.01	0.03	0.860
Error	29	7.20	7.20	0.25		
Total	31	429.81				

Complexity – responses can be affected by multiple causes

- Biological systems are often affected by multiple factors
- Including more than one predictor in a model allows their joint effects to be assessed
- Simple example – volume of lumber is determined by both height of tree and its diameter

Volume of trees is affected by diameter and height





Or, in 3D...

Volume of trees is affected by diameter and height

```
lm(formula = Volume ~ Diam + Height, data = trees)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diam          4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *

Analysis of Variance Table
Response: Volume
          Df Sum Sq Mean Sq  F value  Pr(>F)
Diam       1 7581.8  7581.8 503.1503 < 2e-16 ***
Height     1  102.4   102.4   6.7943 0.01449 *
Residuals 28  421.9    15.1
```

$$\hat{Volume} = -57.98 + 4.7\ Diam + 0.33\ Height$$

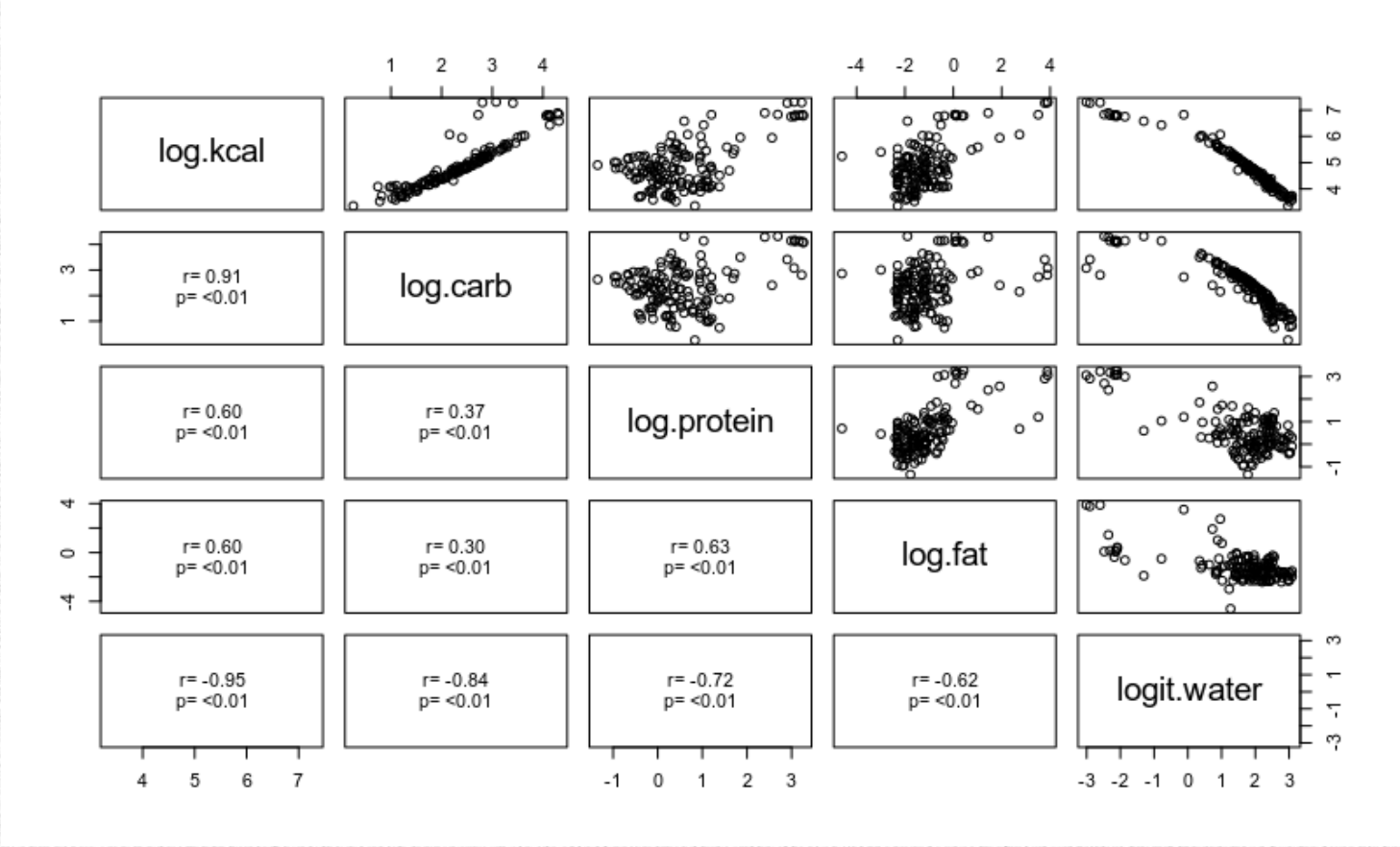
Complication: correlated predictors

- Correlations between predictors = collinearity (multicollinearity)
- Collinearity was responsible for the spurious relationship between AMA and height
- When we know which variable is likely to be spurious we can use the collinearity with other, better predictors to test for spuriousness
- But, we don't always know – collinearity can also mask real relationships
- If two predictors are sufficiently correlated their independent effects on the response can't be told apart
  - Too much shared SS between predictors, not enough independent variation
  - Big standard errors on coefficients (variance inflation)

Collinear predictors can mask real effects

- High correlations between predictors make it difficult to measure their independent effects
- Example: caloric content of food
  - Calories in food is due to digestible macronutrients: carbohydrate, fat, and protein
  - Water does not have any calories, but watery foods are low in all of the macronutrients

The basic patterns



Each predictor, one at a time

```
Response: log.kcal
           Df Sum Sq Mean Sq F value    Pr(>F)
log.carb    1 95.714  95.714  712.02 < 2.2e-16 ***
Residuals 140 18.820   0.134

           Df Sum Sq Mean Sq F value    Pr(>F)
log.fat     1 41.470  41.470  79.461 2.384e-15 ***
Residuals 140 73.064   0.522

            Df Sum Sq Mean Sq F value    Pr(>F)
log.protein  1 40.741  40.741  77.295 4.816e-15 ***
Residuals  140 73.793   0.527

            Df  Sum Sq Mean Sq F value    Pr(>F)
logit.water  1 104.338 104.338  1432.7 < 2.2e-16 ***
Residuals  140  10.196   0.073
```

All the predictors together

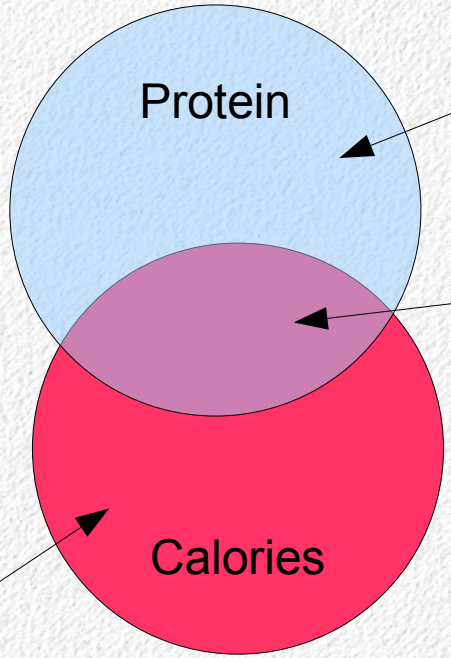
```
Anova Table (Type II tests)

Response: log.kcal
             Sum Sq  Df  F value    Pr(>F)
logit.water 1.1779   1  52.8548  2.49e-11 ***
log.fat     2.4307   1 109.0673 < 2.2e-16 ***
log.carb    4.5828   1 205.6306 < 2.2e-16 ***
log.protein 0.0062   1   0.2785    0.5985
Residuals   3.0532 137
```

What's wrong with this result?

Partitioning variance in a regression

One predictor

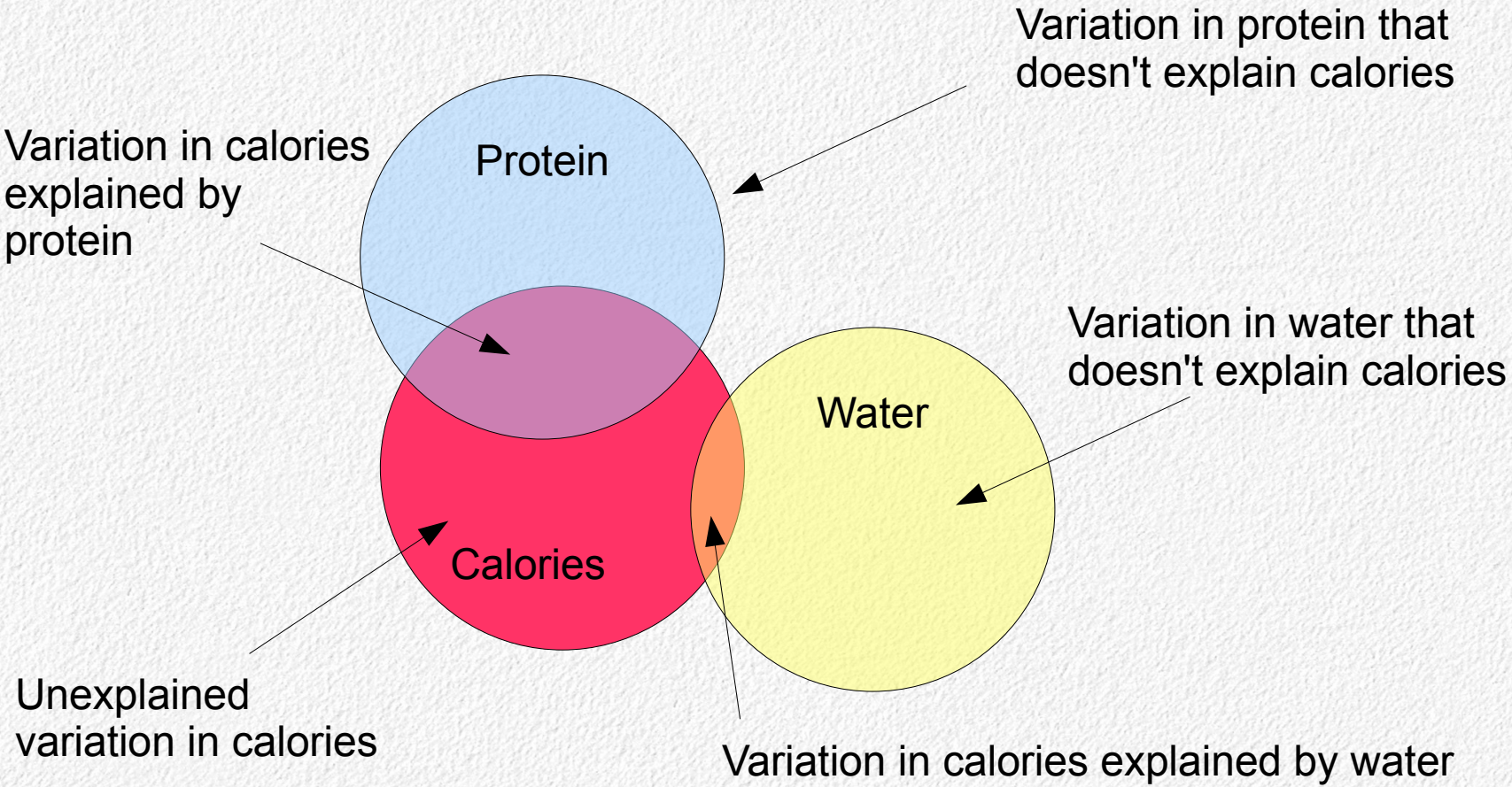


Variation in protein that doesn't explain variation in calories

Variation in protein that explains variation in calories

Unexplained variation in calories

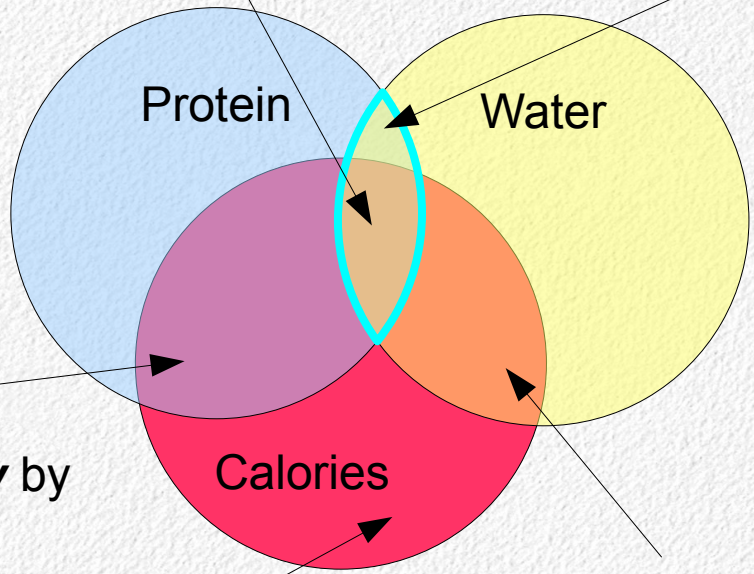
Two predictors that are independent



Two predictors that are correlated

Variation in calories that is ***explained by either protein or water***

Variation in protein and water shared with each other, but that doesn't explain variation in calories



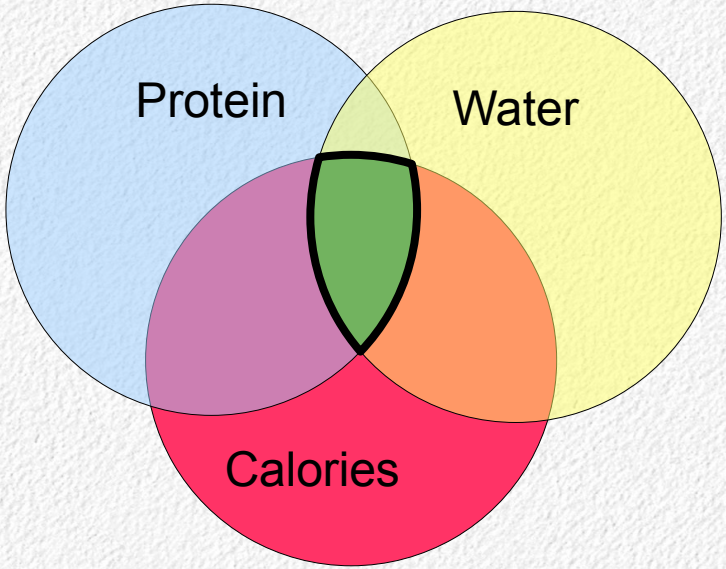
Variation in calories explained ***only*** by protein

Unexplained variation in calories

Variation in calories explained ***only*** by water

What to do with the correlated part?

- Not uniquely attributable to either predictor
- We have choices for how to deal with this:
  - Assign it entirely to one of the predictors –
    Type I sums of squares
  - Remove it from both predictors –
    Type II sums of squares

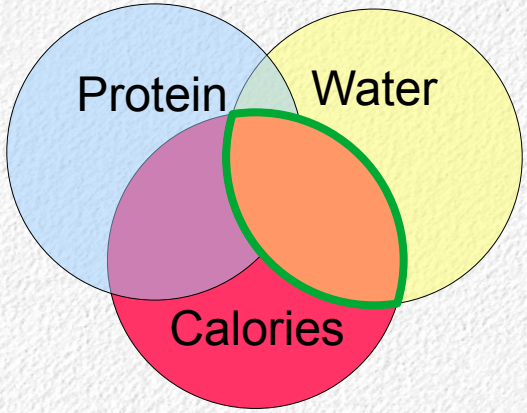


Sequential SS (Type I)

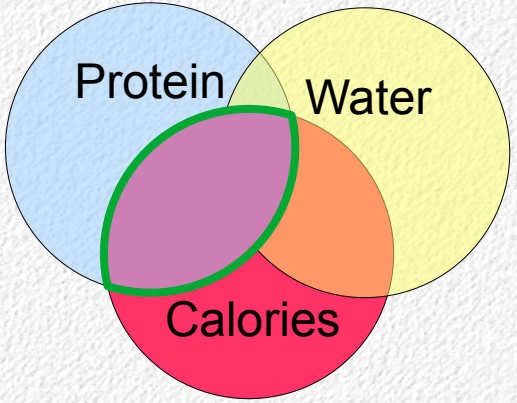
- Predictors added one at a time
  - First predictor is entered, assigned all of the variation in response it explains
  - Second predictor added, only assigned variation it explains in response that isn't already explained by the first predictor (repeat until all entered)
  - Sum of the sequential predictor SS equals to model SS
  - Sum of the predictor SS + residual SS = total SS
- The order that variables are entered affects the results – the ANOVA table p-values won't match the coefficient test p-values, except for the last variable entered

Type I tests with foods data

```
Response: log.kcal
             Df  Sum Sq Mean Sq  F value    Pr(>F)
logit.water   1 104.338 104.338 1781.071 < 2.2e-16
log.protein   1   2.053   2.053   35.041 2.401e-08
Residuals   139   8.143   0.059
```




```
Response: log.kcal
             Df Sum Sq Mean Sq F value    Pr(>F)
log.protein   1 40.741  40.741  695.46 < 2.2e-16
logit.water   1 65.650  65.650 1120.65 < 2.2e-16
Residuals   139  8.143   0.059
```

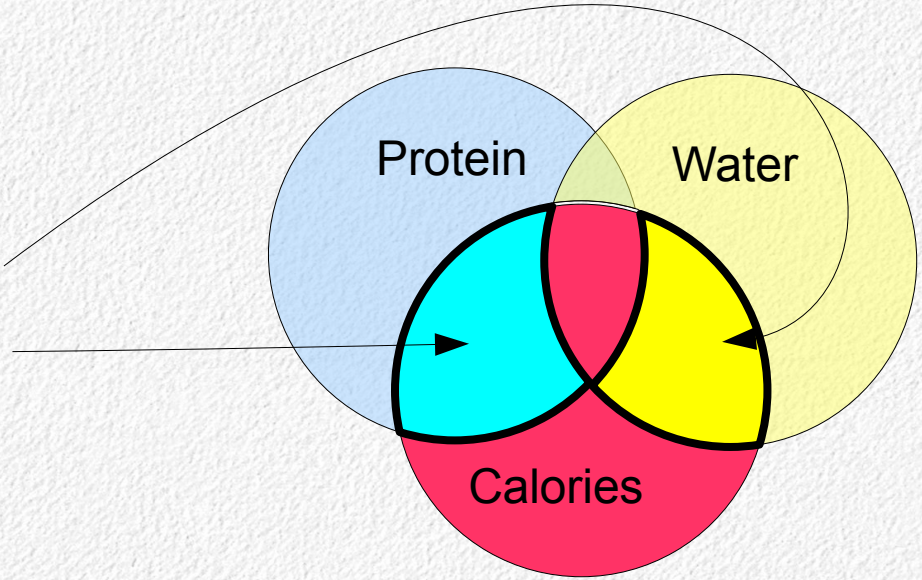


Adjusted SS (Type II)

- Only variation that can be attributed uniquely to each predictor is tested
- Predictors all entered, then one at a time is dropped
  - Loss of model SS after the predictor is dropped is assigned to that predictor
  - Shared variation is not assigned to any predictor (still used for the whole-model omnibus test, and to calculate R^2)
  - Adjusted SS for predictors does not add up to the model SS unless the predictors are perfectly independent
- The order entered doesn't matter
- The p-values in the ANOVA table match the coefficient tests, because both are testing partial relationships

Type II tests for food data

```
Response: log.kcal
             Sum Sq  Df  F value    Pr(>F)
logit.water 65.650   1 1120.650 < 2.2e-16
log.protein  2.053   1   35.041 2.401e-08
Residuals    8.143 139
```



Only the variation explained by a predictor uniquely is tested

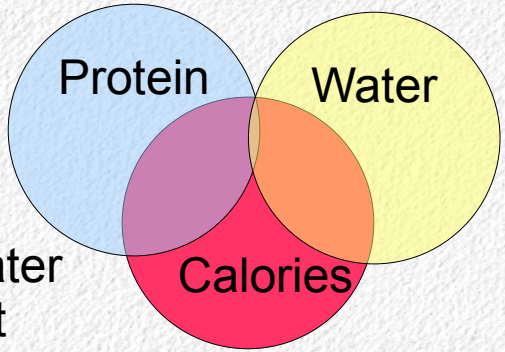
The correlated part is not included in either (so, model + residual SS don’t sum to total)

Note that SS are the same as Type I table for the predictor entered last

Which to use, Type I or Type II SS?

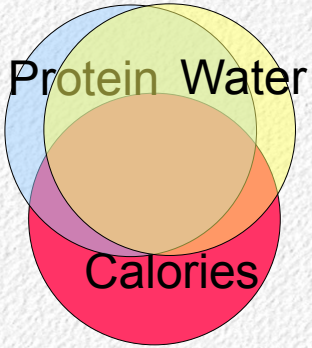
- Depends on the question
- There are ways to design experiments that will make them identical – do this when possible
- If there are correlations between your predictors, and one is a “nuisance”, then either use Type II, or Type I with the nuisance entered first
- If none are known to be nuisances, it's useful to look at the differences between Type I and Type II tests – more on this later

If there is a small correlation



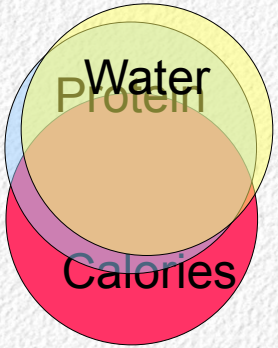
Protein and water both significant with either Type I or II SS

If there is a large correlation, equal amounts of uncorrelated variation that explains calories



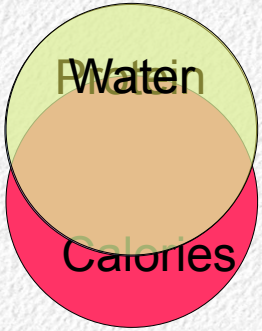
Protein and water both significant for II, first entered much lower p-value than second entered for I

If there is a large correlation, unequal amounts of uncorrelated variation that explains calories



Protein significant for both I and II, water not significant for II, only if entered first for I

If there is a perfect correlation



Can't tell effects of individual predictors, first entered significant for I, neither for II

What's the model?

Which predictors will have positive coefficients? Which negative?

Just one line per predictor, what's happening with the other predictors?

Is it possible the R^2 is higher than 0.95? Why or why not?