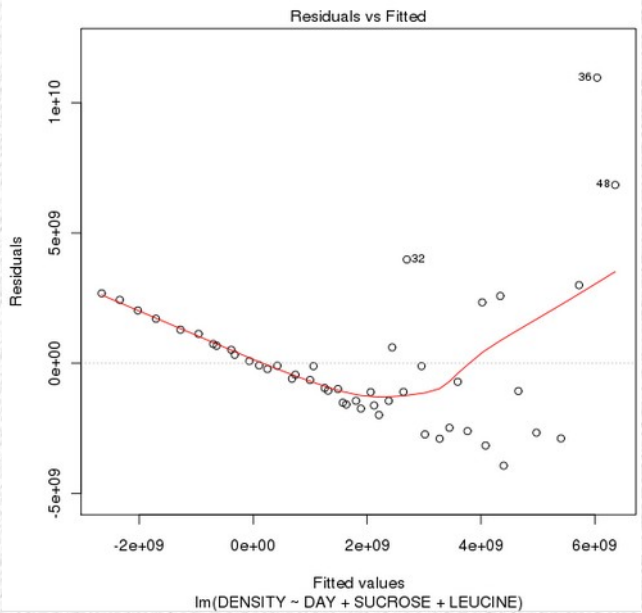
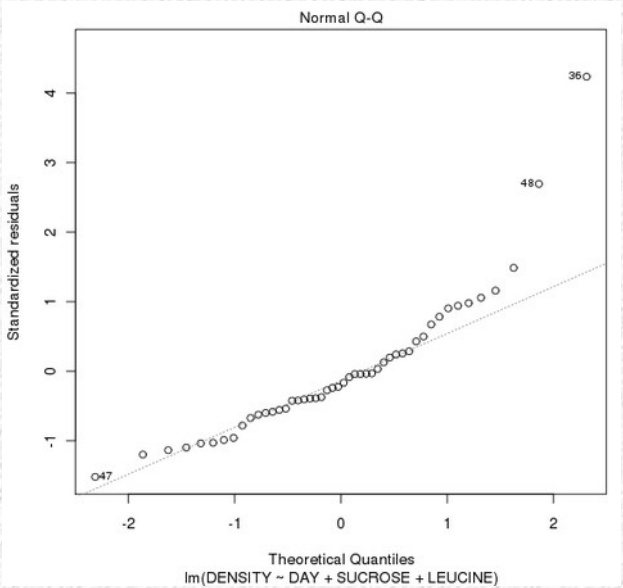
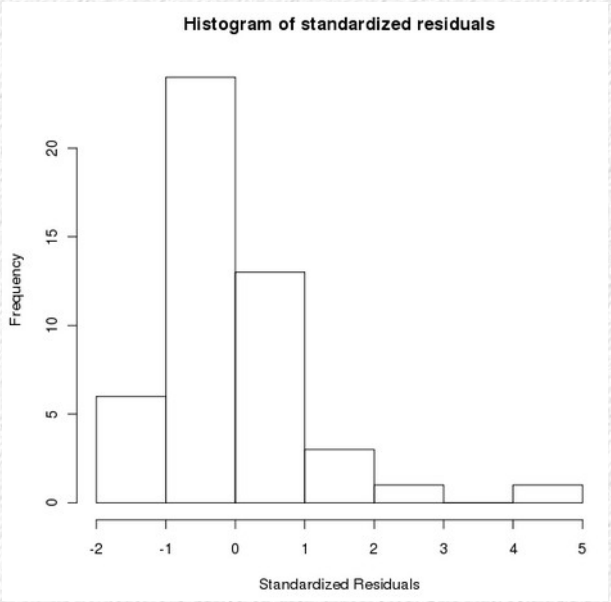
Checking model assumptions







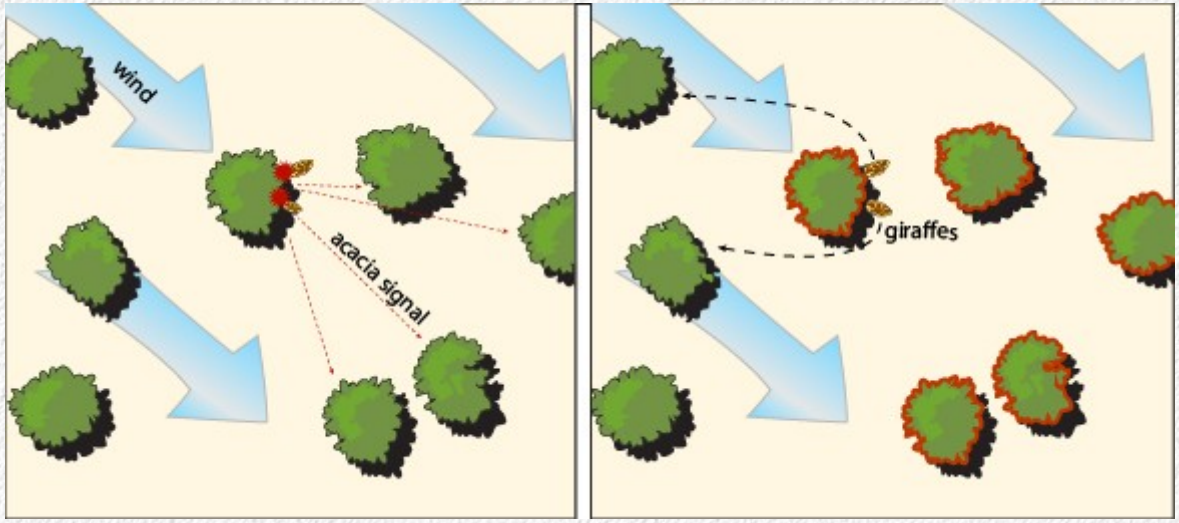
Using models to understand our data

- We use models to interpret our experimental data
  - Coefficients estimate what the effects are
  - p-values tell us if the effects are non-random
  - Variation explained by the model tells us how strong the relationships are
- We are responsible for making sure the models we use are appropriate for our data
- Any model we use has limits – if we use it improperly we can't expect good results
  - We can only see the effects the model looks for
  - p-values are only accurate if our data have the properties assumed by the model
- The properties we need our data to have for a model to work properly are called model **assumptions**
- We often speak of assumptions as though the are a judgment about the data, but it's really a judgment about whether a model is appropriate for the data

Where assumptions come from

- General – conditions that need to be true for sample data to give reliable answers about populations
  - Independent observations – need multiple, distinct measurements of response
  - Random sampling – samples must be representative of the population
- Specific – conditions derived from the structure of the model we use, how the p-values are calculated
  - Linearity – straight line relationship between numeric predictors and numeric responses
  - Equal variances
  - Normality

The trees are talking to each other!



Why is this a problem?

How to solve it?

Independence of measured responses

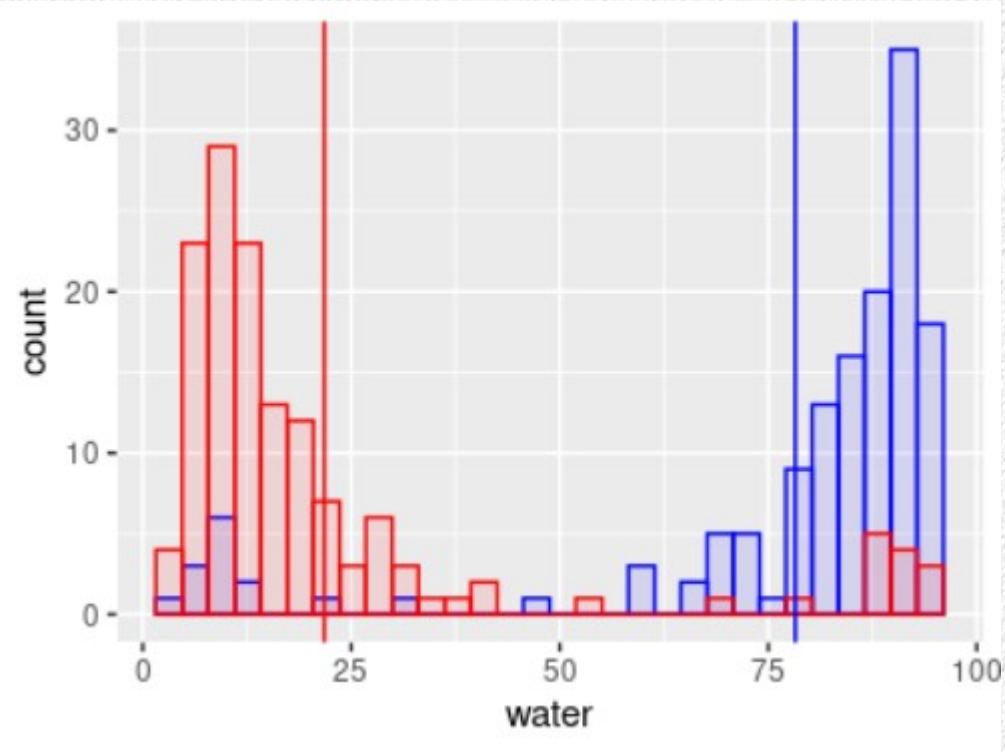
- Statistical definition of independence of events (i.e. responses):

  If occurrence of one event has no effect on the probability of another event occurring, they are independent events

- Can fail to have independent events if:
  - Experimental subjects influence one another
  - Some uncontrolled, unmeasured variable is influencing observations
  - Use of repeated measurements
- Not the same as independence of **variables**
  - Purpose of our experiments is to test if a response **variable** is affected by a predictor
  - If a predictor affects the response, then the response variable is not independent of the predictor variable
  - Independence of predictor and response is not assumed – detecting dependency between predictor and response variables is the reason to do the study in the first place

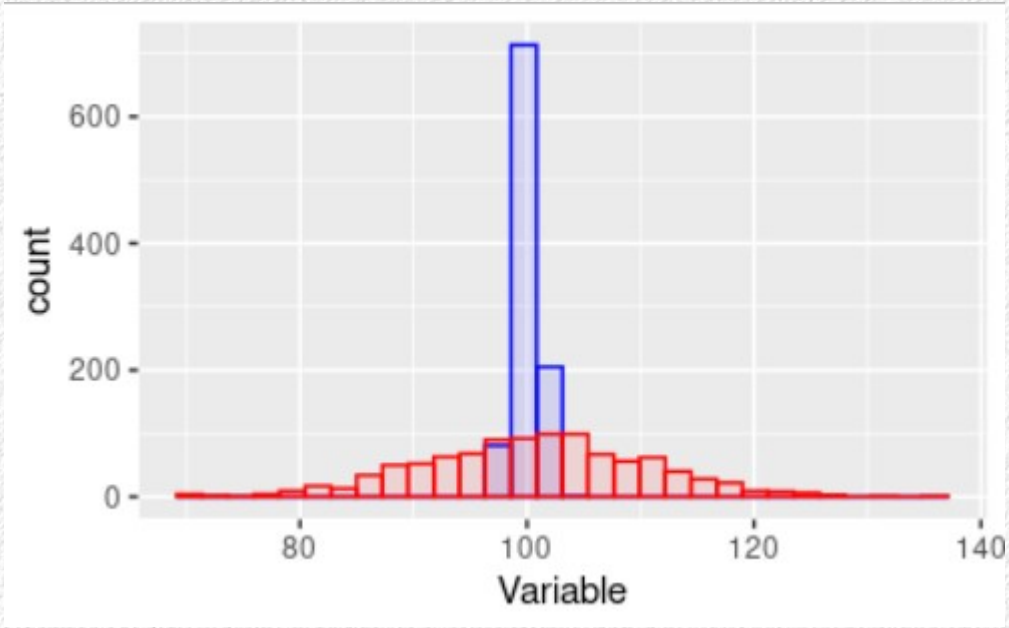
Normality

- There is nothing wrong with data that are not normally distributed
- But, GLM is based on the assumption that data are distributed normally around means (or around predicted values)
  - p-values assume normality
  - Non-normal distributions that are the same may be okay with large n
  - Non-normal distributions that are different can be problematic even with large n, because the mean becomes a misleading measure of typical response
- We should avoid analyzing non-normal data with GLM



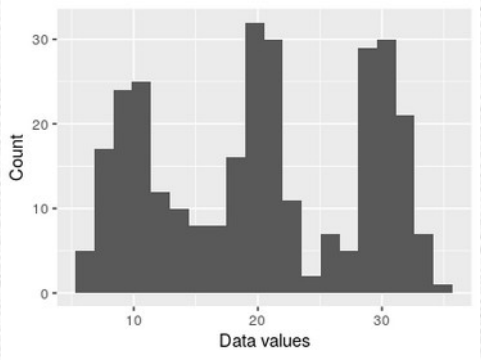
Equal variances

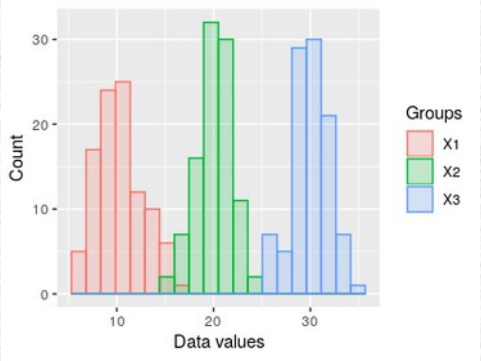
- With different variances, but the same mean, differences in sample means will be large more often
- If we don’t account for this, we would have more false positives than we should
- Having equal variances avoids this problem



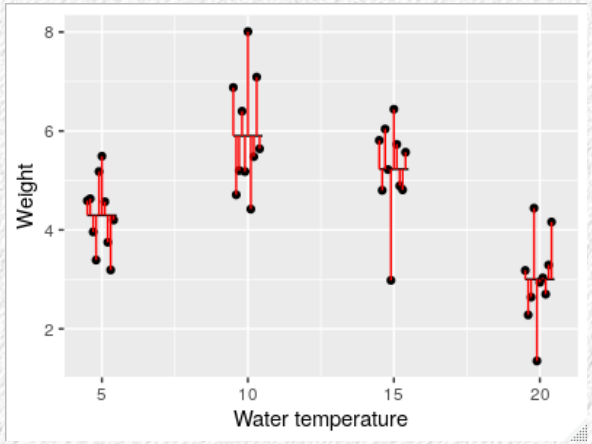
We have been testing normality wrong

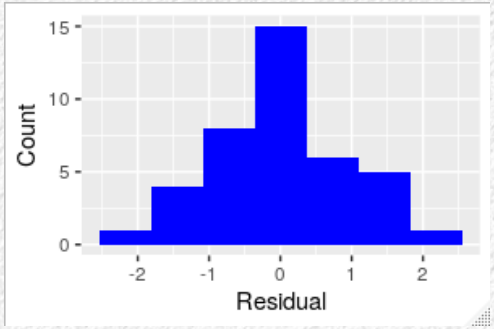
- Testing for normality is subject to an unfortunate tradeoff:
  - Normality is a bigger issue for small sample sizes (why? Let's see...)
  - Assumption tests have Ho: the assumption is met
  - So, detecting departures from normality is hardest with small sample sizes
- To test normality, we have been splitting the data into groups and test separately (why?)
- Splitting the data by groups reduces sample size, makes it less likely we will detect departures from normality when they matter most
- Istead, we will start using residuals to test assumptions

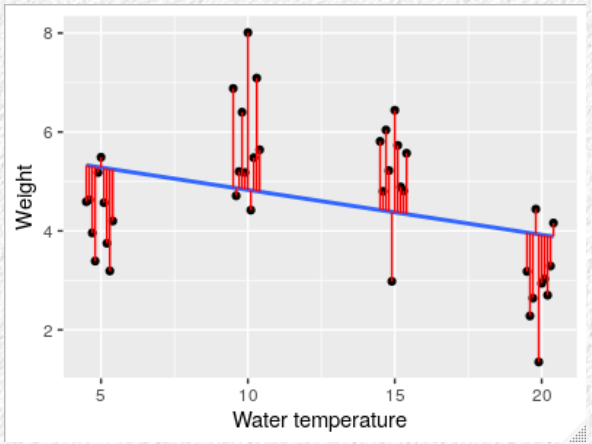


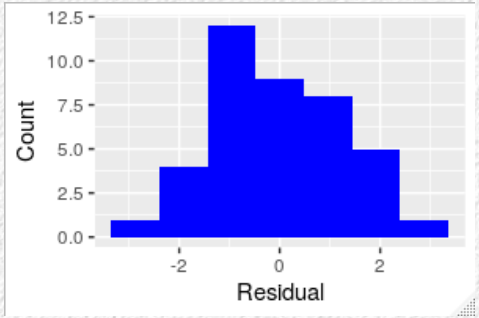


Distribution of residuals depends on the model









Assumptions checked by inspecting residuals

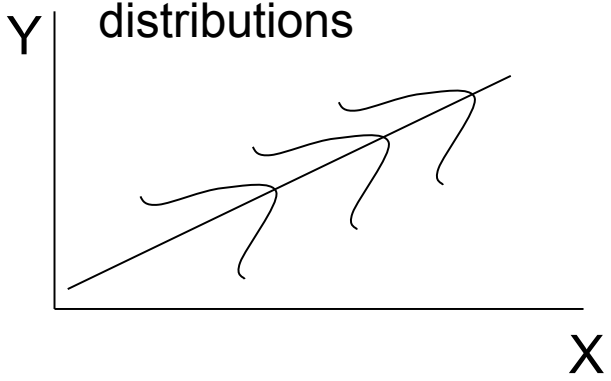
- We expect residuals to represent random variation
  - Unpatterned
  - Independent
- GLM requires them to be normally distributed
- We will (primarily) use graphical tools to assess:
  - How well the model fits the data
  - If the data have the distribution needed to use the model to interpret our experiment

Assumptions of GLM's

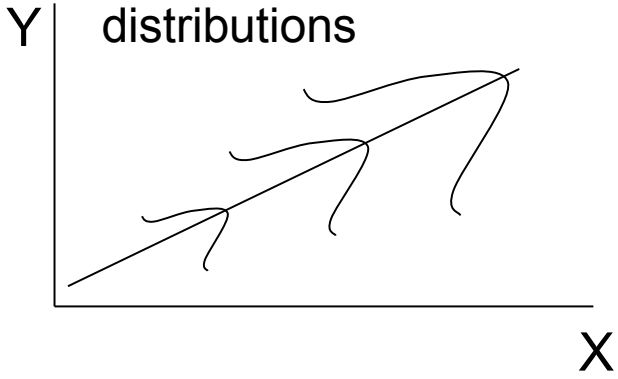
- General assumption (independence of errors, random sampling)
- Model-based assumptions
  - Normality = residuals are normally distributed around the predicted values from a model
  - Homogeneity = the variance in residuals is the same around all predicted values from a model
  - Linearity = there is a straight line relationship between response and any numeric predictor used

Homogeneity of variances/normality for residuals along regression lines

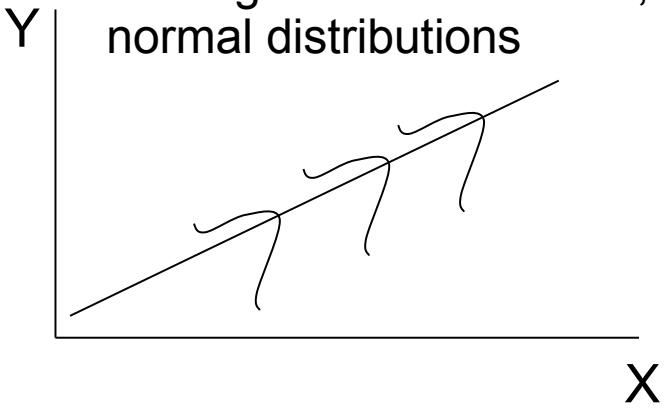
Homogeneous variances, normal distributions



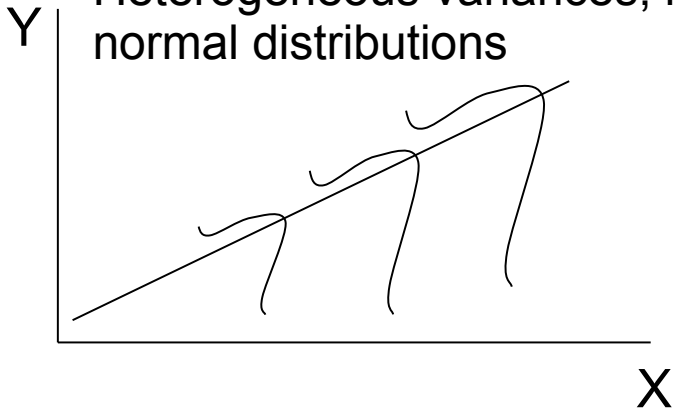
Heterogeneous variances, normal distributions



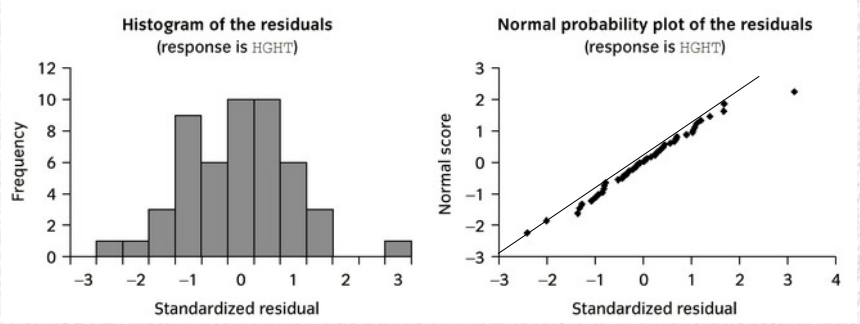
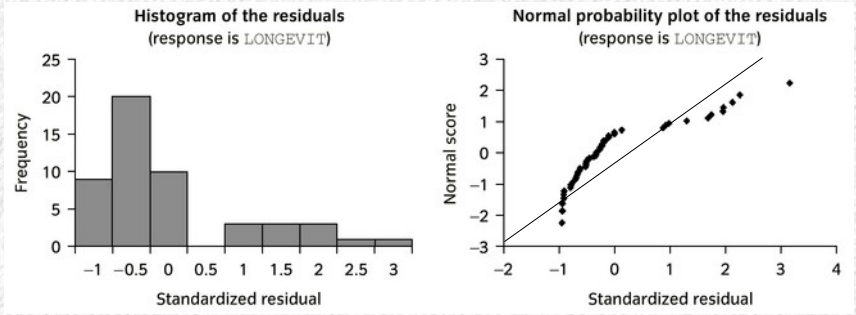
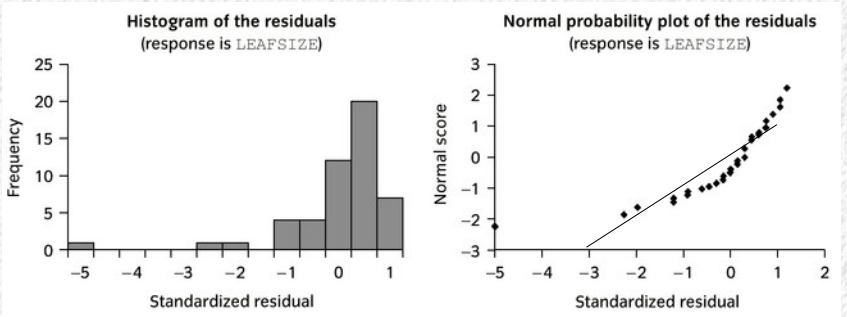
Homogeneous variances, non-normal distributions



Heterogeneous variances, non-normal distributions

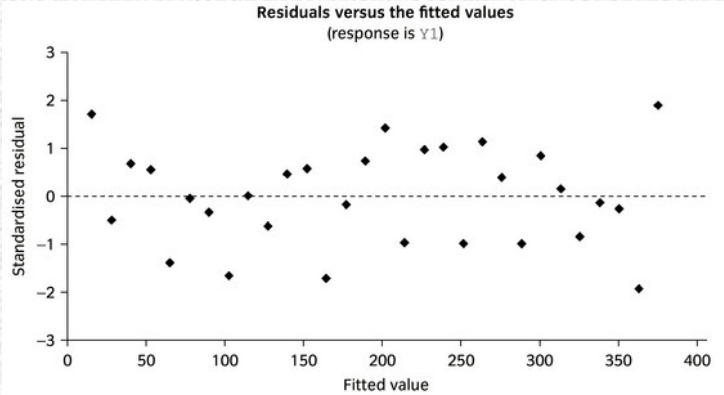


Normality

	Histogram of residuals	Normal probability plots
Normality assumption met		
Right-skewed residuals		
Left-skewed residuals		

HOV, linearity, independence of data points

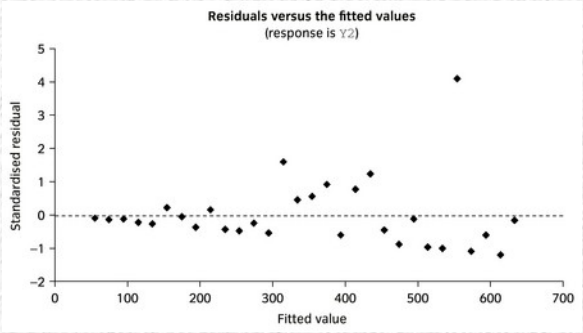
Good



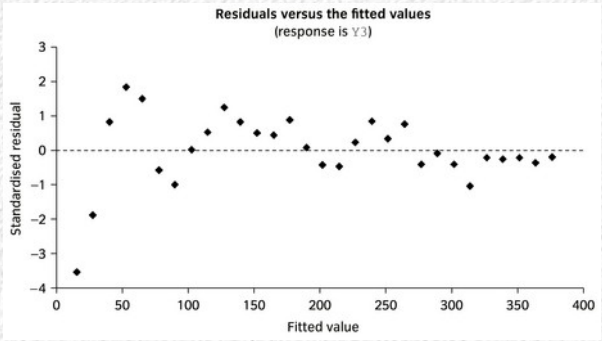
Residual vs. fitted value plots

Not Good

Variance increases with predicted value



Variance decreases with predicted value

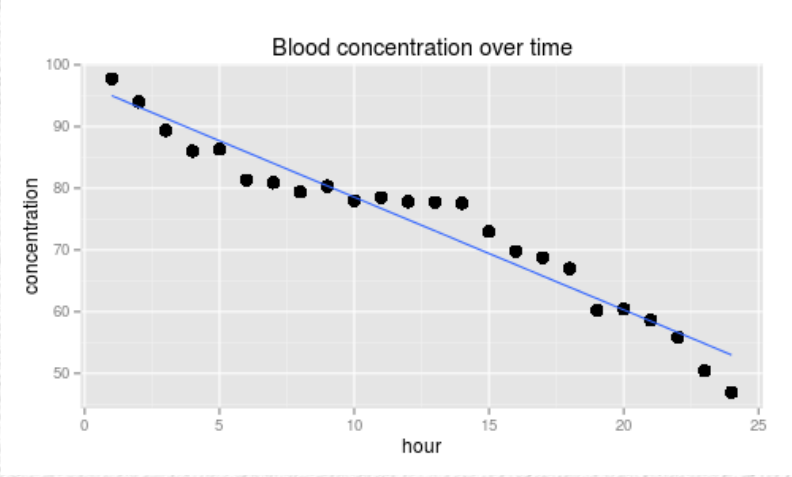


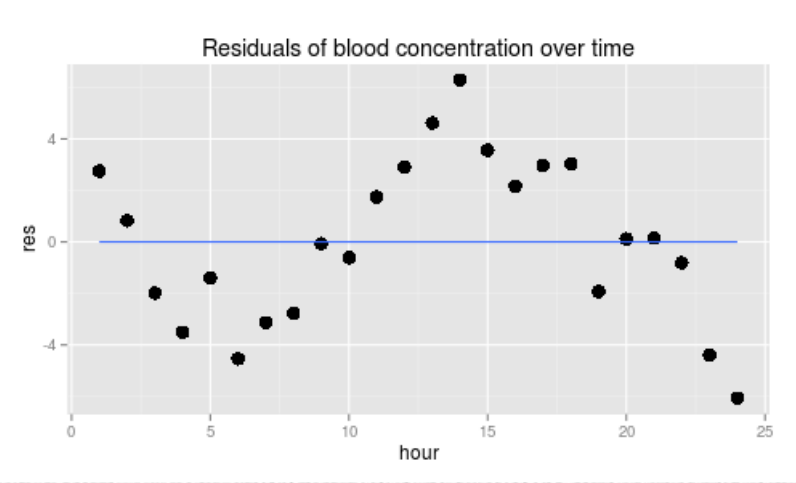
What if you don’t meet GLM assumptions?

- There are several possible treatments:
  - Add a variable
  - Add an interaction between variables
  - Apply a transformation
- If none of those work, use a different analysis

Example: adding a variable

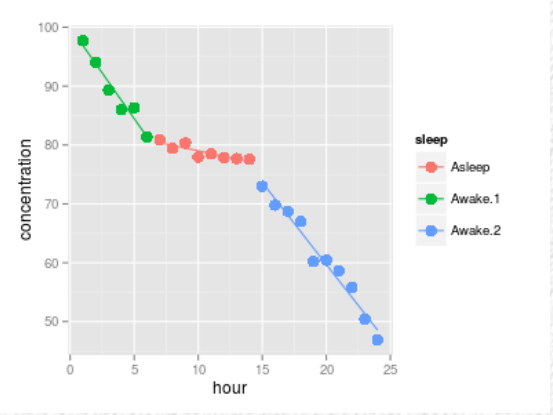
- (contrived) data on blood concentration of a compound each hour after it was administered
- The data seems to be changing slope in a predictable way, but the line isn't capturing this
- Produces a pattern in the residuals (they are “temporally autocorrelated”)

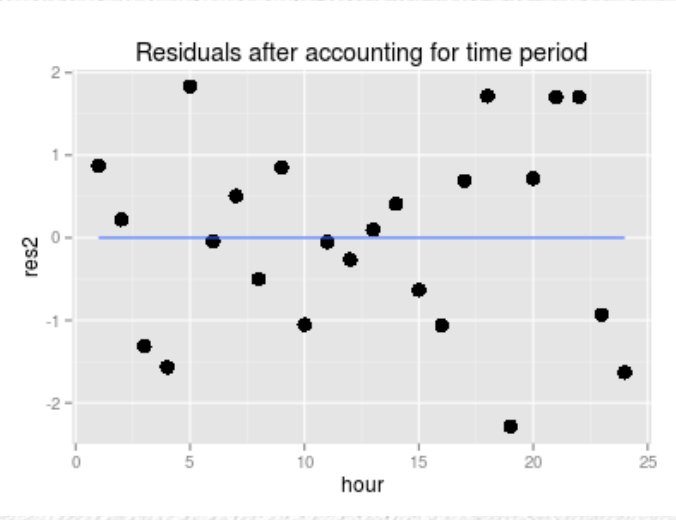




Accounting for the dependency

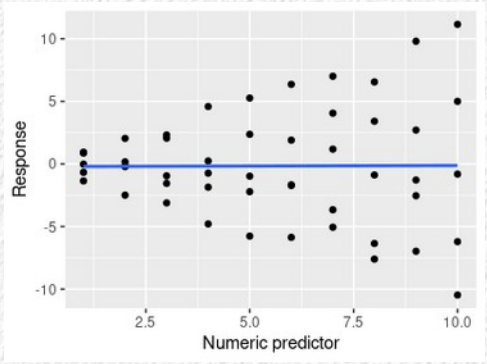
- Design your study to avoid dependencies of errors if possible
- If not possible to avoid, can model the dependency
  - Include a variable (sleep) that records if the subject is awake or asleep
  - Include sleep status as a predictor
  - The dependency due to this variable is thus accounted for
  - Residuals become unpatterned → independent

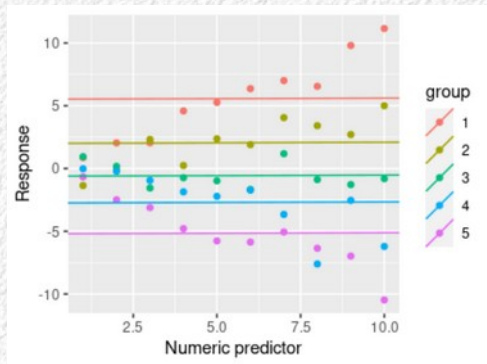


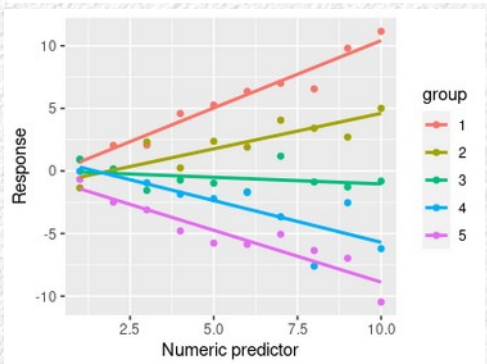


Adding interactions to fix HOV problems

- Lack of HOV can be due to an unmeasured variable, or an interaction that isn't accounted for
- Example here – increased variance from low to high values of numeric predictor
  - Including the group variable helps some
  - Including an interaction between numeric predictor and group accounts for the pattern
- What's left is HOV







Model criticism

- Since the residuals depend on the model, we can't test our assumptions first
  - Have to fit a model, then test the residuals
- The **model criticism** process:
  1) Fit a model to the data
  2) **Inspect**/test the distribution of residuals
  3) Add interactions, additional variables, or apply a transformation
  4) Repeat as needed until you meet model assumptions
- Once you have a model that fits the data, only interpret that model

Example: bacterial growth experiment

- Example: test of effects of leucine, sucrose levels on bacterial growth
  - Response = bacterial density
  - Predictors = leucine level (3), sucrose level (4), day of sample (4)
  - Factorial design used → all possible combinations (complete), equal numbers (balanced)
- The simplest model for these data would be:

  Density ~ Day + Sucrose + Leucine

- How well does the model fit the data?

Initial model

BOX 9.8 Analysing bacterial growth without interactions

General Linear Model

Word equation: DENSITY = DAY + SUCROSE + LEUCINE

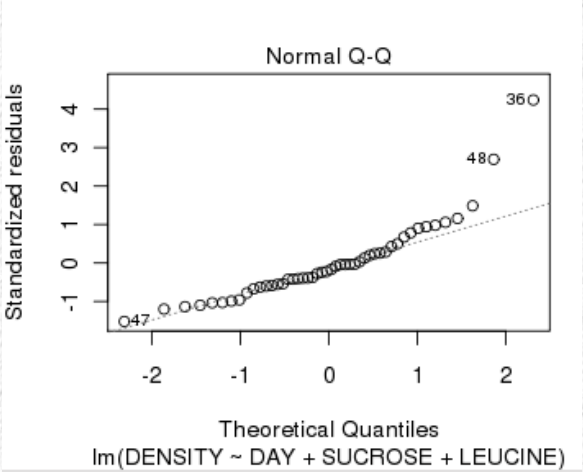
DAY, SUCROSE and LEUCINE are categorical

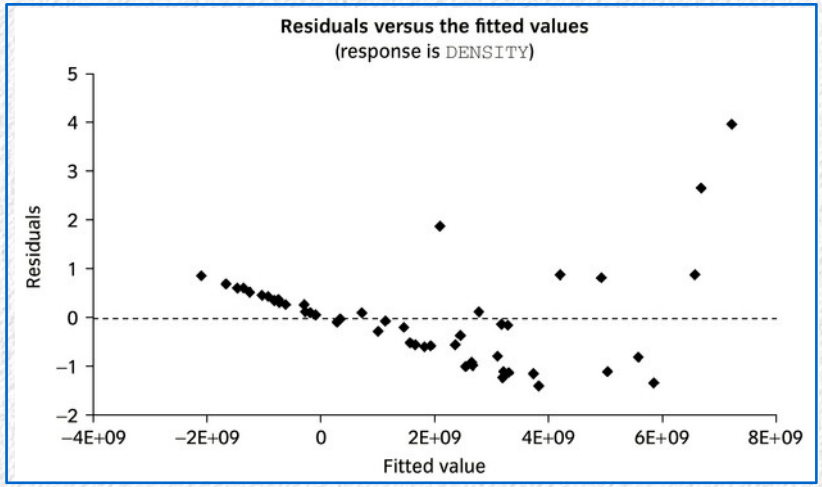
Analysis of variance table for DENSITY, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
DAY	3	1.1570E+19	1.1570E+19	3.8566E+18	0.52	0.674
SUCROSE	3	1.1895E+20	1.1895E+20	3.9651E+19	5.31	0.004
LEUCINE	2	1.4762E+20	1.4762E+20	7.3811E+19	9.88	0.000
Error	39	2.9136E+20	2.9136E+20	7.4709E+18		
Total	47	5.6951E+20				

Fit?

First step – add an interaction





Model with an interaction between sucrose and leucine

BOX 9.9 Reanalysis of bacterial growth, including the interaction

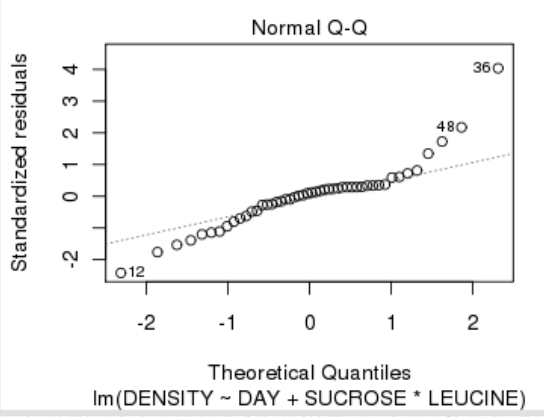
General Linear Model

Word equation: DENSITY = DAY + SUCROSE + LEUCINE + SUCROSE * LEUCINE

DAY, SUCROSE and LEUCINE are categorical

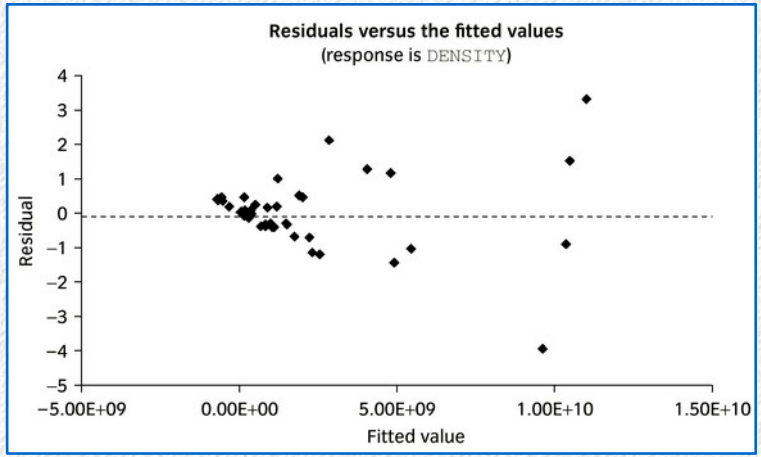
Analysis of variance table for DENSITY, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
DAY	3	1.1570E+19	1.1570E+19	3.8566E+18	0.81	0.496
SUCROSE	3	1.1895E+20	1.1895E+20	3.9651E+19	8.36	0.000
LEUCINE	2	1.4762E+20	1.4762E+20	7.3811E+19	15.56	0.000
SUCROSE * LEUCINE	6	1.3479E+20	1.3479E+20	2.2464E+19	4.73	0.001
Error	33	1.5658E+20	1.5658E+20	4.7447E+18		
Total	47	5.6951E+20				



Better fit – more linear, but still heterogeneous variance

Next, try a transformation

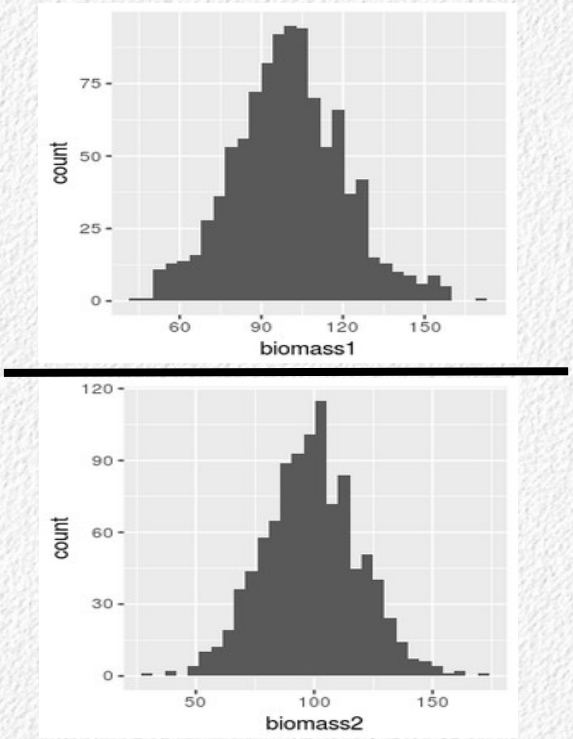


Right-skewed data

Right-skewed variables are common in biology

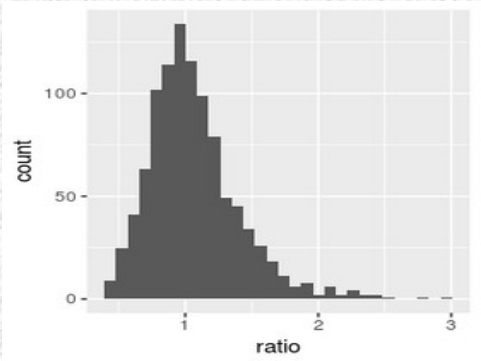
Result when there are basements = minimum possible values (usually 0)

True for dimensions, ratios of numbers



=

Skewed ratio of two normal variables



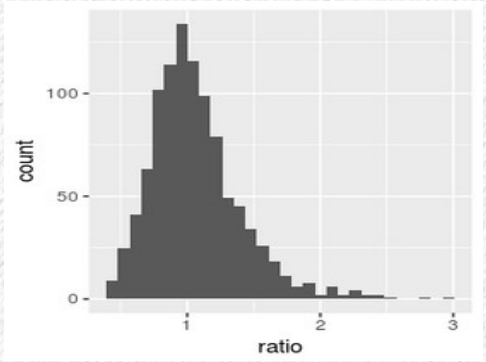
To can use a mathematical function that changes the scale of the variable to make it normally distributed

Called a transformation

Has to shrink the upper tail, expand the lower tail = non-linear change, change in the relative spacing between data values

Common transformations

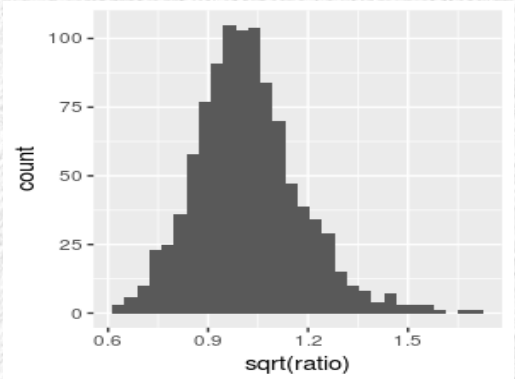
- Right-skewed distributions – in order of increasing strength
  - Square root
  - Log
  - Negative inverse
- Right-skewed variables often have variances that increase with the mean, so transformation treats both normality and HOV

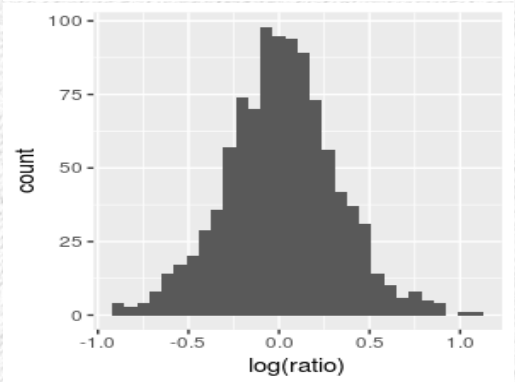


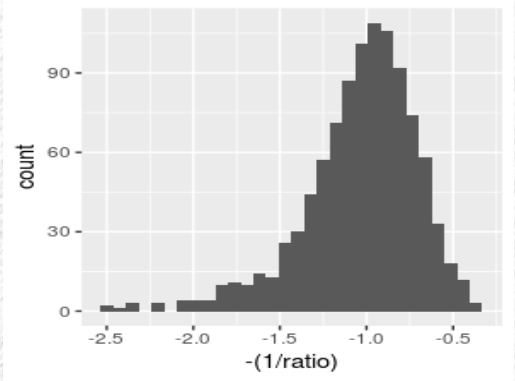
√

ln

-1/x

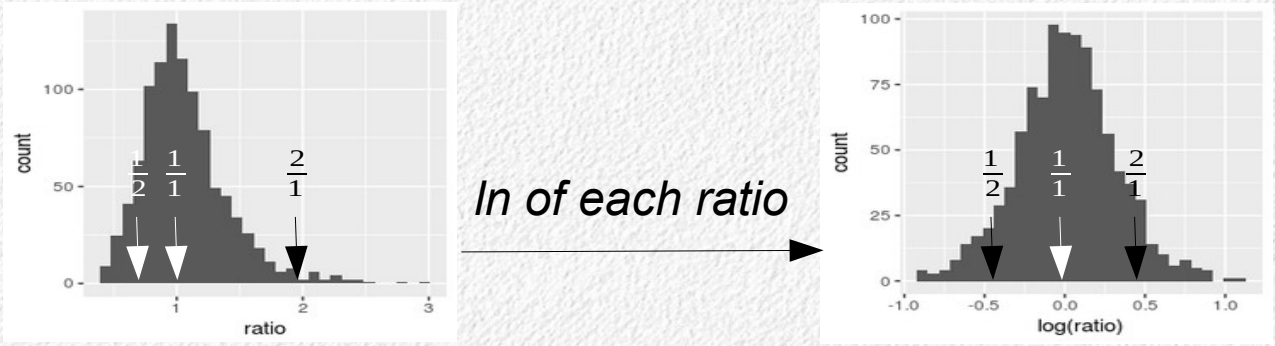




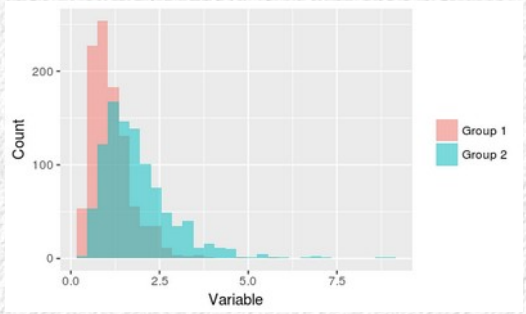


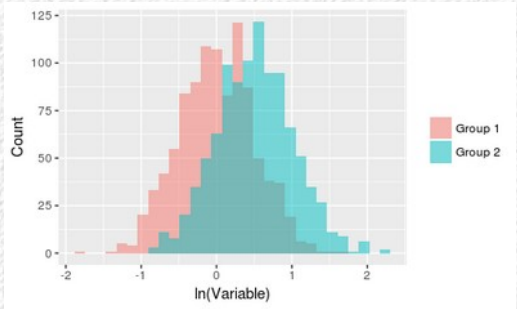
Log transformation can improve normality and HOV

Log scale compresses large numbers, expands small numbers



When an increase in mean → increase in variance, log transformation often makes variances equal

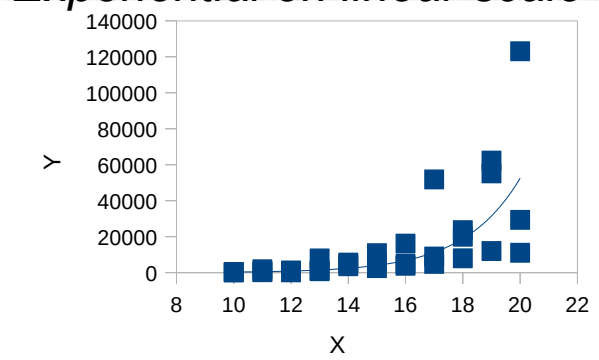




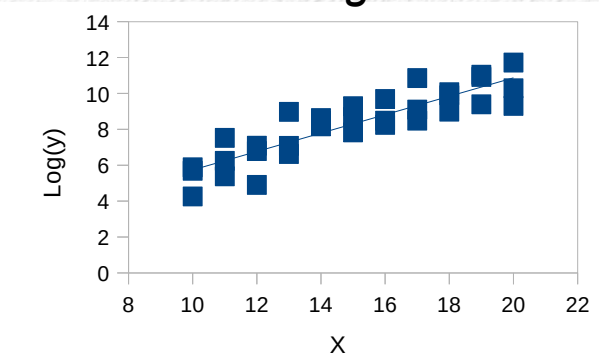
Log transforming to improve linearity

- We may also do a log transformation to address a lack of linearity in the data
- Exponential relationships become linear after log transforming the response variable

Exponential on linear scale



Linear on log scale



Log transform the dependent variable (bacterial density)

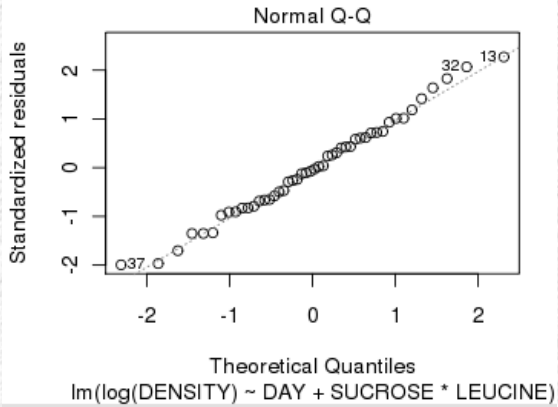
BOX 9.10 Reanalysis of bacterial growth with transformation

General Linear Model

Word equation: LOGDEN = DAY + SUCROSE + LEUCINE + SUCROSE * LEUCINE
DAY, SUCROSE and LEUCINE are categorical

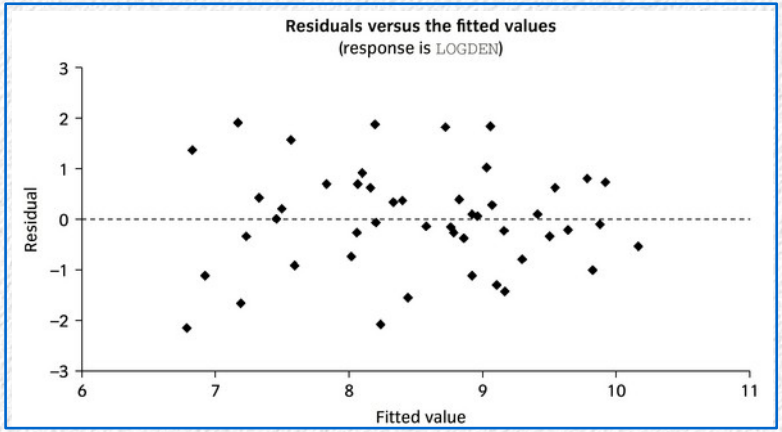
Analysis of variance table for LOGDEN, using Adjusted SS for tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
DAY	3	1.0461	1.0461	0.3487	1.38	0.265
SUCROSE	3	20.8387	20.8387	6.9462	27.55	0.000
LEUCINE	2	15.1785	15.1785	7.5892	30.10	0.000
SUCROSE * LEUCINE	6	1.1489	1.1489	0.1915	0.76	0.607
Error	33	8.3204	8.3204	0.2521		
Total	47	46.5326				



Good fit – linear, homogeneous variances

Interpret this one!



But, transformation changes your analysis

- For illustration, focus on a comparison of leucine levels 1 and 3
- Means of ln(density) are:
  Leucine 1 = 18.11
  Leucine 3 = 21.26
- Difference between them is 3.15
- What does this mean?

Back-transformation

- To convert from log scale back to the data units, we need to **back-transform** the log-scale values
  - Apply the inverse function
  - For logs, this is the exp function = raise the base of the logs to the power of the mean
- $e^{18.11}$ = exp(18.11) = 73,294,784
  $e^{21.26}$ = exp(21.26) = 1,710,411,805
- Arithmetic means of density are:
  Leucine 1 = 397,412,688, Leucine 2 = 4,313,368,750
- The values used in the GLM are not arithmetic means – so, what are they?

Arithmetic means on a log scale are geometric means on a linear scale

Arithmetic mean on a log scale

$$\bar{x} = \frac{\sum \log(x)_i}{n}$$

Geometric mean on a log scale

$$GM_x = \sqrt[n]{\prod x_i}$$

Difference between arithmetic means on a log scale

$$21.26 - 18.11 = 3.15$$

(what the GLM uses)

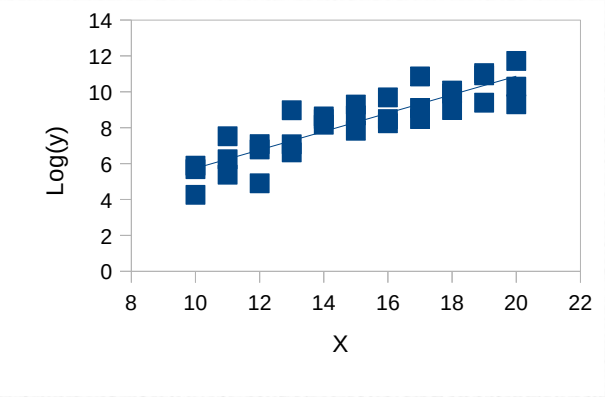
Ratio of GM's on a linear scale

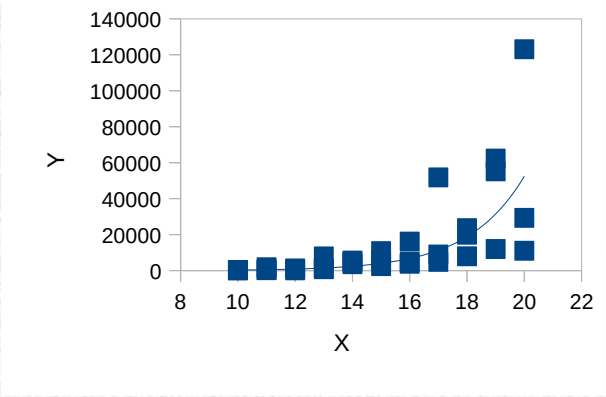
$$e^{21.26-18.11} = e^{21.26} / e^{18.11} = 23.34$$

$$\frac{GM_{\text{Leucine 1}}}{GM_{\text{Leucine 3}}} = \frac{1{,}710{,}411{,}805}{73{,}294{,}784} = 23.34$$

(Interpret this – geometric mean density at Leucine level 1 is 23.34 times bigger than at Leucine level 2)

Linear on a log scale, exponential on a linear scale





$$\hat{\log(y)} = \log(k) + mx$$

$$\hat{y} = k\,10^{mx}$$

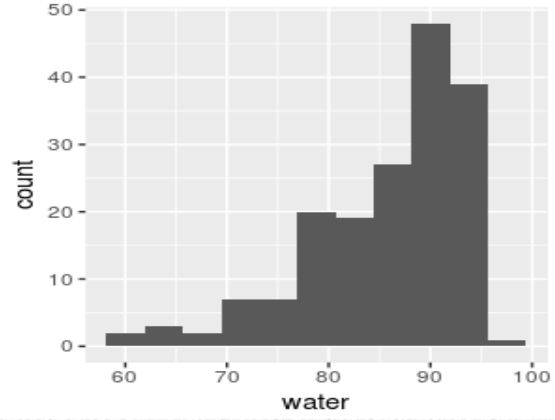
If log(y) has a straight line relationship with x, then y has an **exponential** relationship with x

Meaning, y is related to x as an exponent of a base

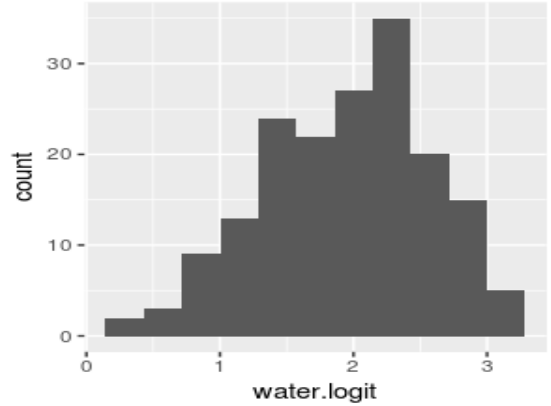
Other data types require different transformations

- Example: proportions and percentages
- Data can be either right or left skewed:
  - Basement of 0
  - Ceiling of 1
- Data are fairly bell-shaped when mean is near 0.5

Before transformation



After transformation

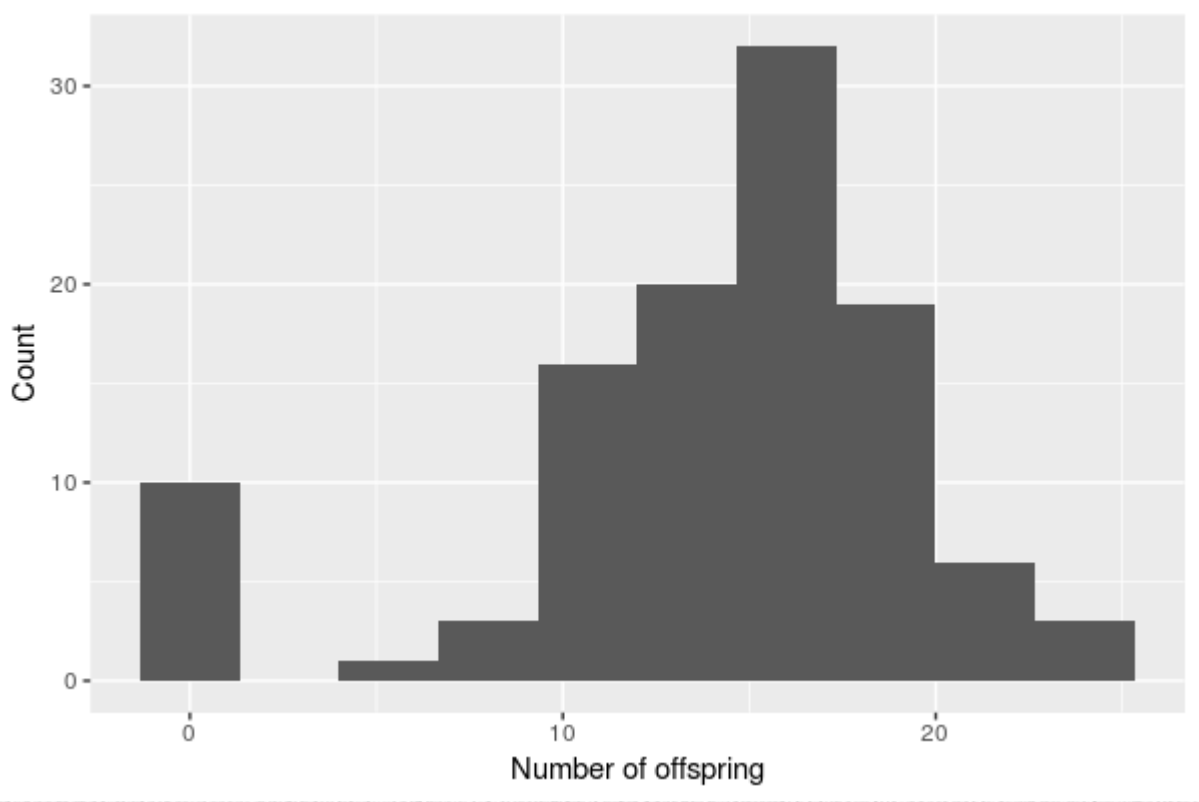


- Can use a **logit transformation**, which is the log odds ratio:

$$logit(p) = \ln\left(\frac{p}{(1-p)}\right)$$

- Best if done in a “generalized linear model” that uses the logit as a “link” function – beyond the scope

A distribution that transformation won't fix



Lots of repeated data values cause problems

Any transformation will transform all to the same value

For a distribution like this, may be necessary to use another approach, such as a randomization test, or a “zero-inflated” model

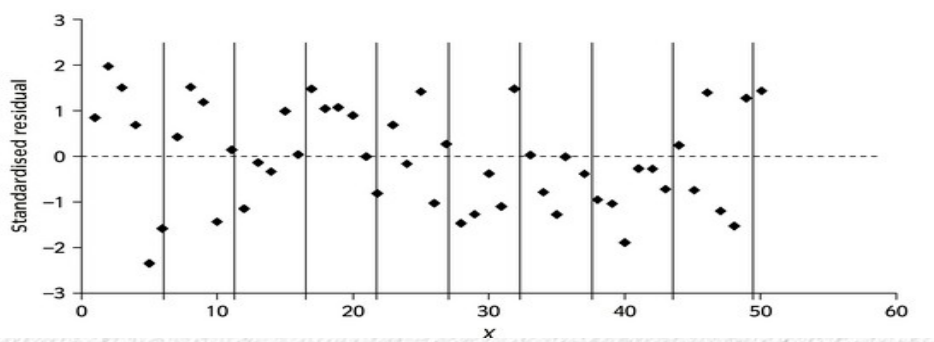
Why rely on graphical tools?

- There are quantitative tests of these assumptions
- Problem is...
  - The larger your sample size, the greater power to detect even small violations of assumptions

    but....
  - the larger your sample size, the less these violations of assumptions matter
- Quantitative tests of violations of assumptions are often no improvement over careful, thoughtful inspection of graphs of residuals

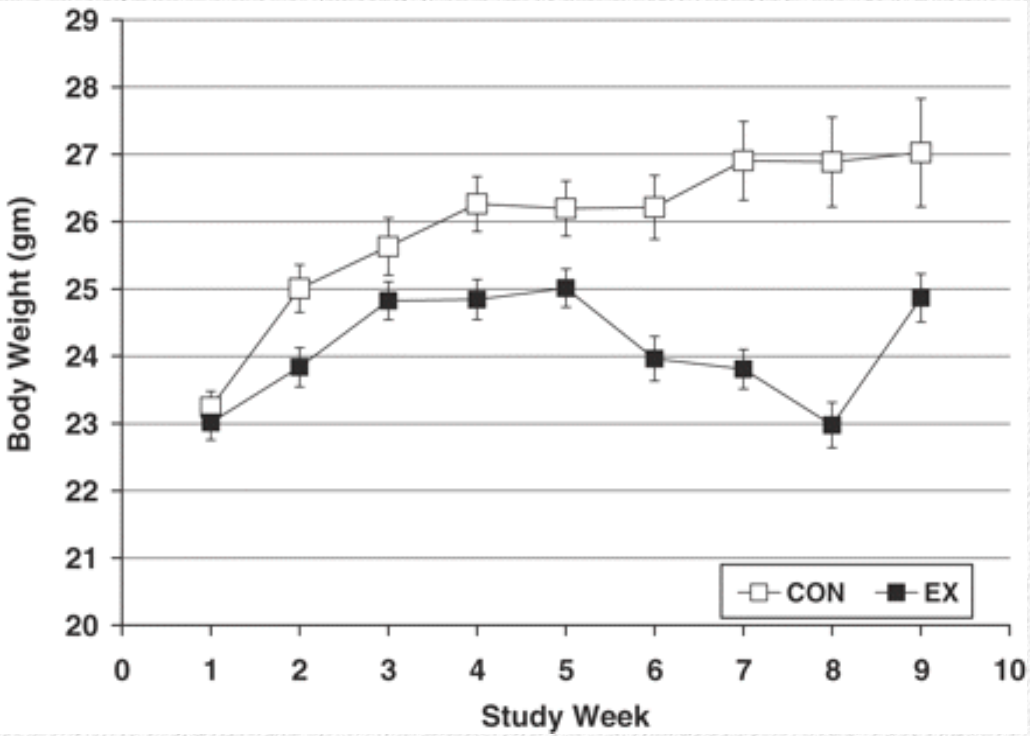
Practical advice

- Don't be too picky
  - GLM's are robust to minor violations of assumptions
  - They become more robust the larger the sample size
  - If the graphical methods look good, you shouldn't worry



- Focus on a small number of transformations that work in most cases
- If violation of assumptions is severe, use alternative methods (non-parametric tests, randomization tests)
- If numeric covariates are used, try transforming them as well to fix nonlinearities

What's the model?



CON = control

EX = mice allowed to voluntarily wheel run