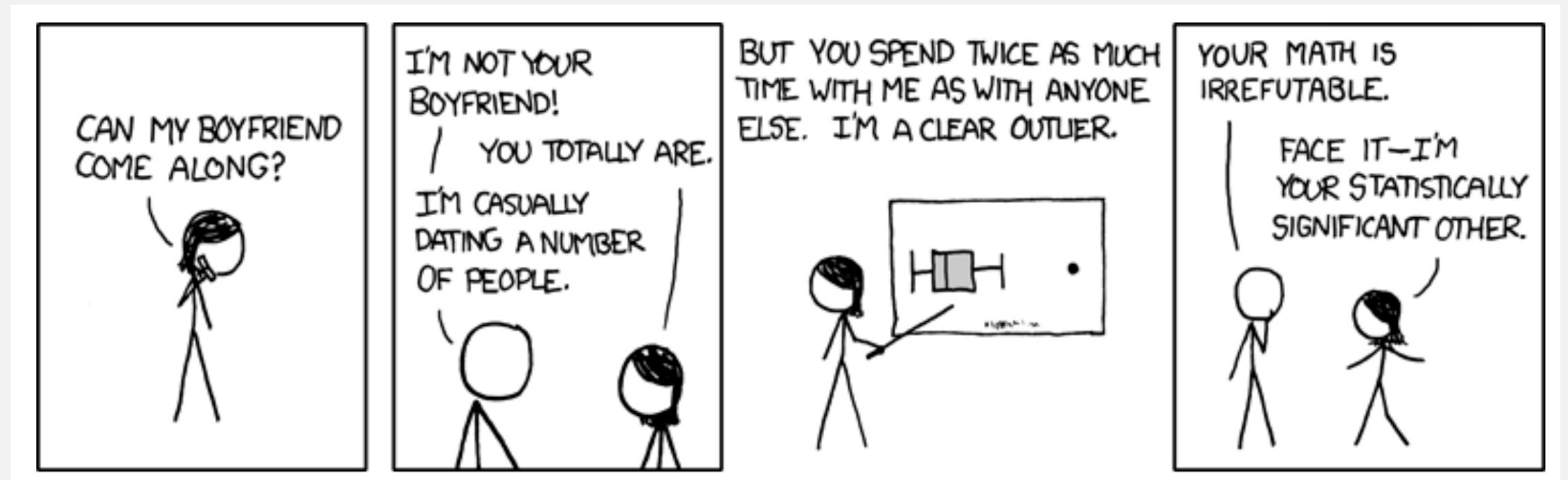


# Binary Response Variables, Random vs Fixed Effects, and Outliers

## Biology 683

### Lecture GLM2

Heath Blackmon



# Other kinds of regression

**Logistic regression** allows us to fit a binary response variable (absent/present; alive/dead) with one or more categorical or continuous predictor variables.

**Poisson regression** allows us to fit a response variable that is Poisson distributed (number of extinctions in a unit of time, number of colonies per plate, (number of occurrences for rare events)) with one or more categorical or continuous predictor variables.

```
fit.logi <- glm(obs ~ pred2 , family="binomial")
```

```
fit.pois <- glm(obs ~ pred2, family="poisson")
```

# GLM and LM function in R

- The GLM and LM function in R takes equations that can be described with the following operators

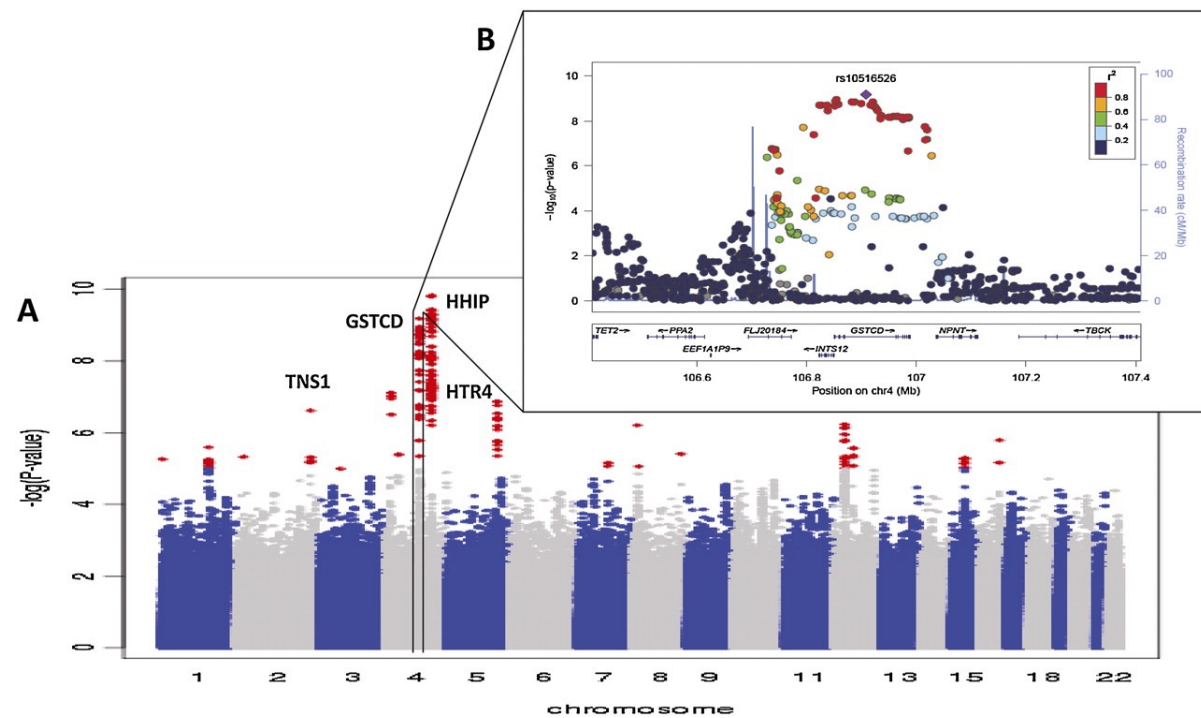
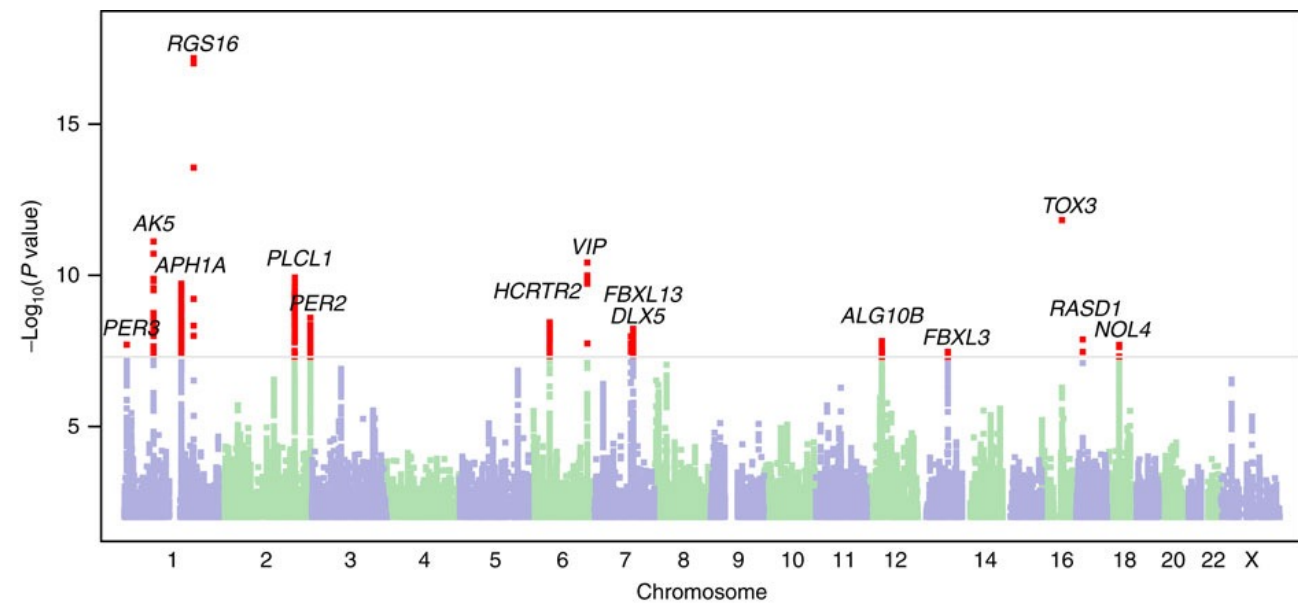
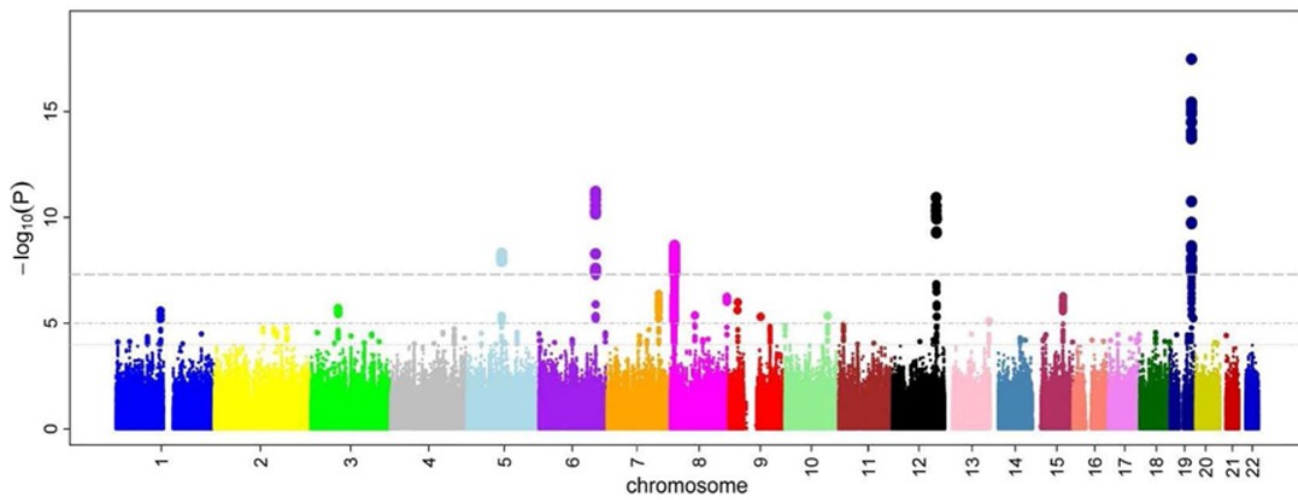
+	+X include this variable
:	X:Z include the interaction between these variables
*	X*Y include these variables and the interactions between them
^	(X + Z + W)^3 include these variables and all interactions up to three way

# R versus the math implied

$$\text{glm}(y \sim X + W) \quad y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i$$

$$\text{glm}(y \sim X * W) \quad y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \epsilon_i$$

# GWAS



# GWAS

## **GWAS: Genome wide association study.**

The goal of GWAS is to determine what genes have alleles that are responsible for a trait of interest. The trait can be any measurable trait in any organism that you wish to study. For instance, a disease in humans, an economically important trait of a crop or domestic animal, an adaptation like a certain color pattern in birds, etc.

# GWAS – Discrete condition (often disease)

cases

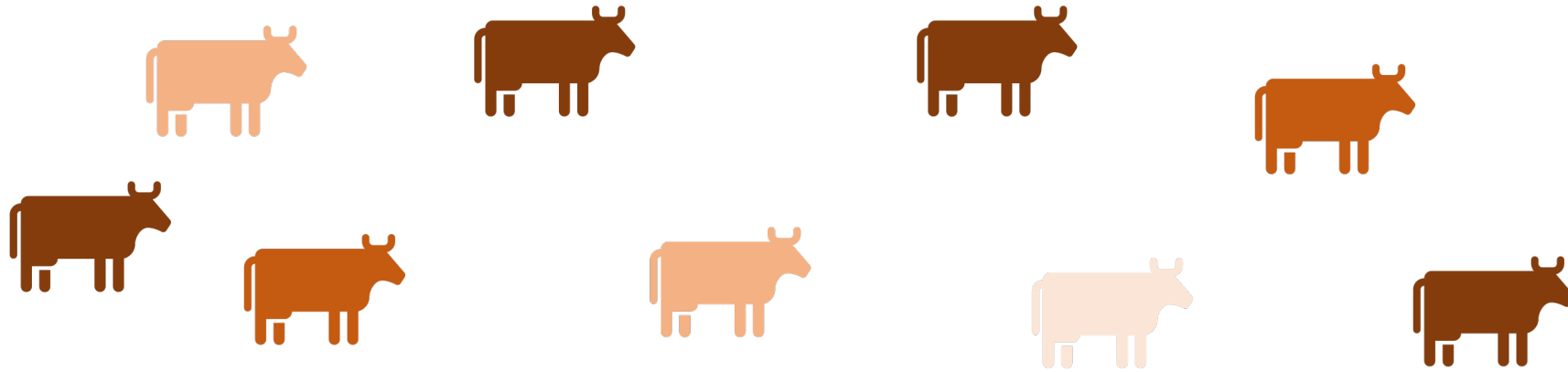


controls



```
>case 1
CATACTACTACTGAACGTTTGCTCCTGCTactatctctctctctctctctctctctctctctCATGC
>case 2
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>case 3
CATACTACTACTGAACGTTTGCTCCTGCTactatctctctctctctctctctctctctctctCATGC
>control 1
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>control 2
CATACTACTACTGAACGTTTGCTCCTGCTactatctctctctctctctctctctctctctctCATGC
>control 3
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
```

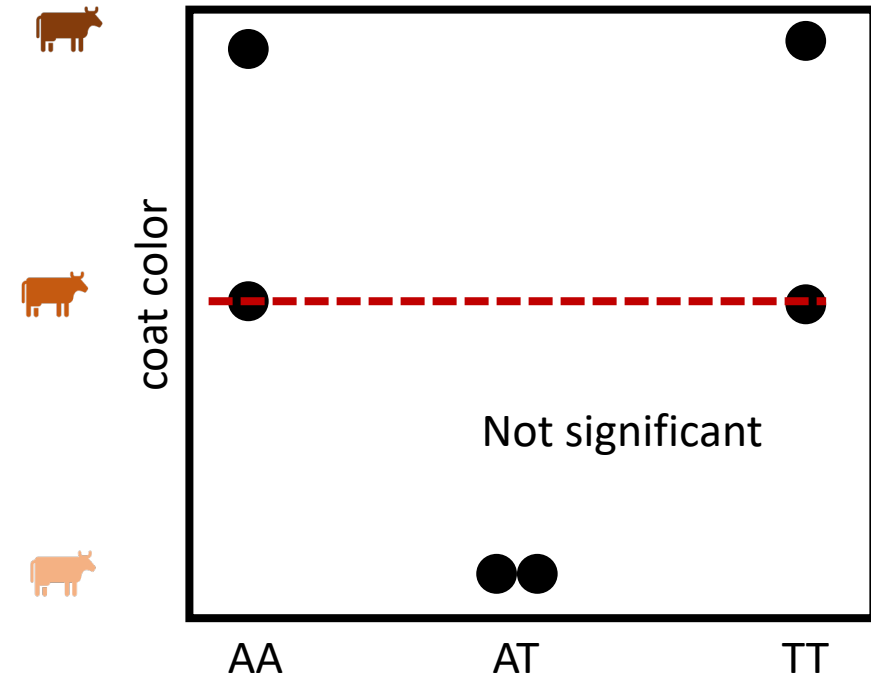
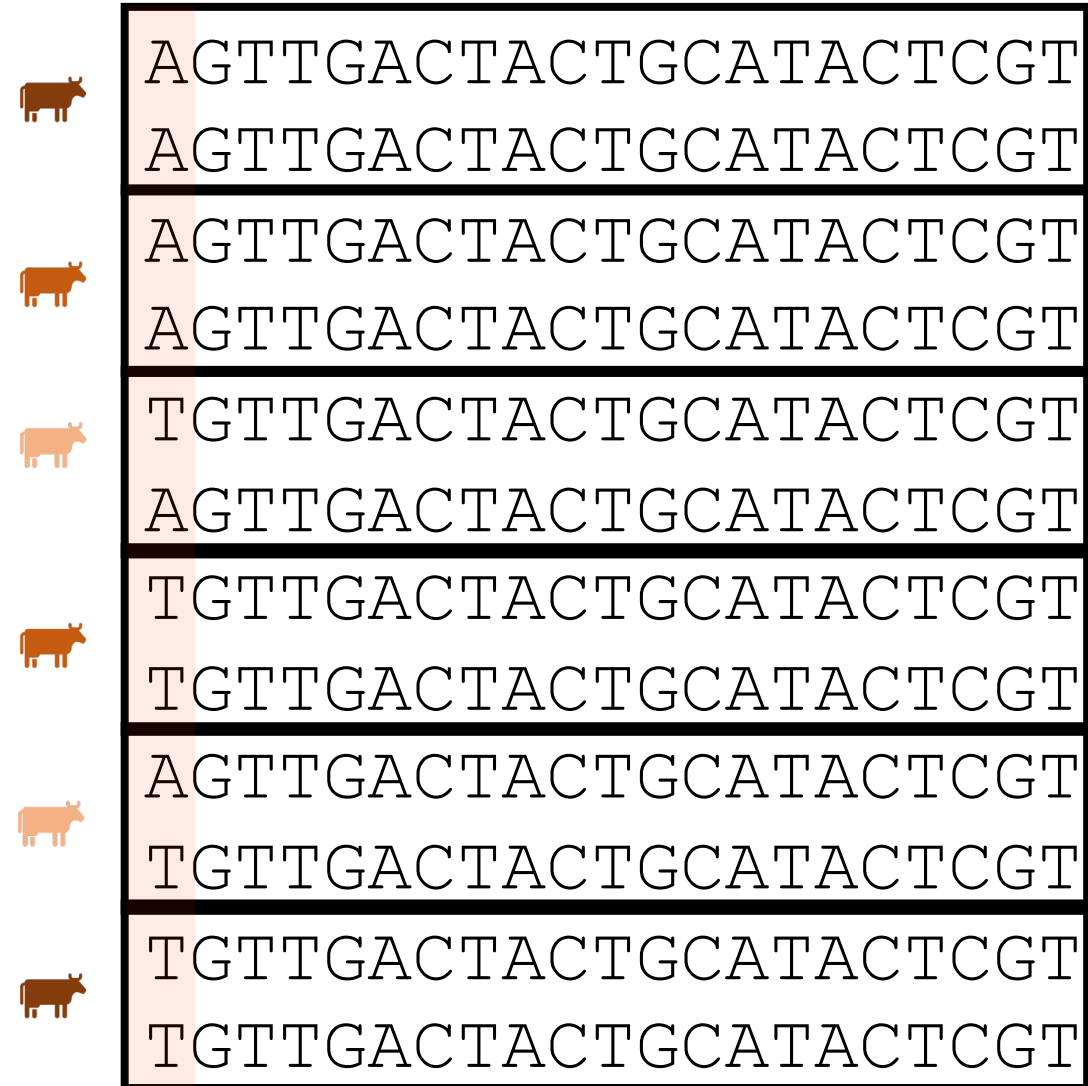
# GWAS – Continuous trait









```
>sample 1
CATACTACTACTGAACGTTTGCTCCTGCTactatctctctctctctctctctctctctctctctCATGC
>sample 2
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>sample 3
CATACTACTACTGAACGTTTGCTCCTGCTactatctctctctctctctctctctctctctctctCATGC
>sample 4
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
>sample 5
CATACTACTACTGAACGTTTGCTCCTGCTactatctctctctctctctctctctctctctctctCATGC
>sample 6
AGTTGACTACTGCATACTCGTGCTAGCTGACTGTCGTACGTACGTAGCTAGTGATCGATCGATGCTAGCTA
```

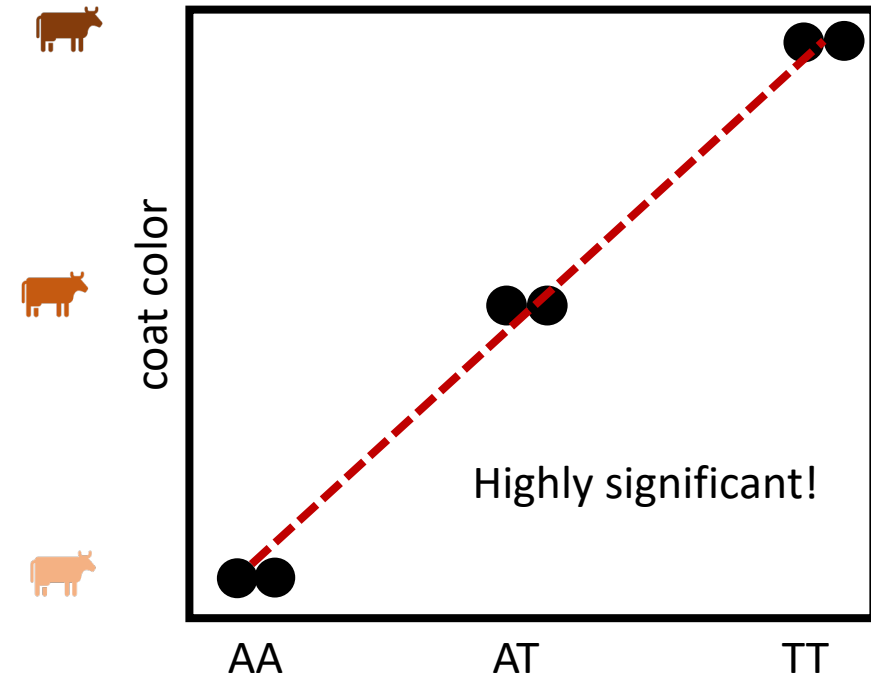


# GWAS – Continuous trait



# GWAS – Continuous trait

	AGTTGACTACTGCATACTCGT
	AGTTGACTACTGCATACTCGT
	AGTTGACTACTGCATACTCGT
	AGTAGACTACTGCATACTCGT
	TGTAGACTACTGCATACTCGT
	AGTAGACTACTGCATACTCGT
	TGTAGACTACTGCATACTCGT
	TGTTGACTACTGCATACTCGT
	AGTAGACTACTGCATACTCGT
	TGTAGACTACTGCATACTCGT
	TGTTGACTACTGCATACTCGT
	TGTTGACTACTGCATACTCGT



# GWAS - Problems

**What is the problem with doing this across the whole genome?**

**Multiple tests lead to more false positives!**

- 1) require a higher level of significance  $5 \times 10^{-8}$
- 2) only look at the very most significant
- 3) lots of more complicated approaches too!

**What is one of the most basic requirements of almost all statistical tests?**

**Tests normally assume independence of the data points!**

- 1) Samples from a population will be related to each other due to ancestry (trees!)
- 2) DNA sequencing is not done equally in all groups of people (western samples are usually over represented)

# GWAS - Problems

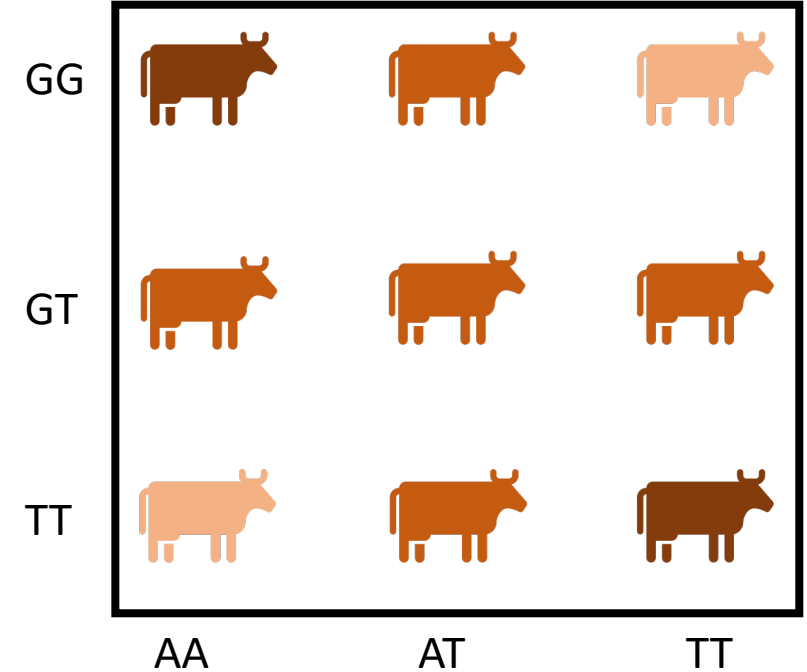
## What is epistasis?

**It is the case where the impact of a genotype at one locus depend on the genotype at another locus!**

To find this type of effect you would need to look at all pairs of genotypes. How many tests would we need to do then?

About 6 orders of magnitude more tests would be required.

Humans have around 4,000,000 sites that are variable like this which equates to 7,999,998,000,000



# GWAS - Problems

**What about the environment?**

**Many diseases have a strong environmental component (heart disease, diabetes, cancer, etc.)**

If these are left out of the study often what is discovered is actually genetic variation that happens to coincide with environmental factors?

If a disease is more common in Europeans than Africans or Asians but it is because of a lifestyle characteristic any genetic variation that is common in Europeans but rare in Africans and Asians could appear associated with the disease.

# Mixed models

Mixed models are models that include fixed and random effects.

Fixed effects can be repeated by other researchers. These are the variables that you are interested in studying.

Random effects are usually nuisance parameters. These are variables that other researchers cannot replicate and you are not interested in inferring anything about them.

# Mixed models

Fixed effects are the variables whose impact we wish to determine

- Characteristics of the media or habitat
- Experimental treatments
- Age groups
- Time points
- Mutant genotypes

Conclusions that you reach are only applicable to the groups or treatments you include in the study

# Example of fixed effects

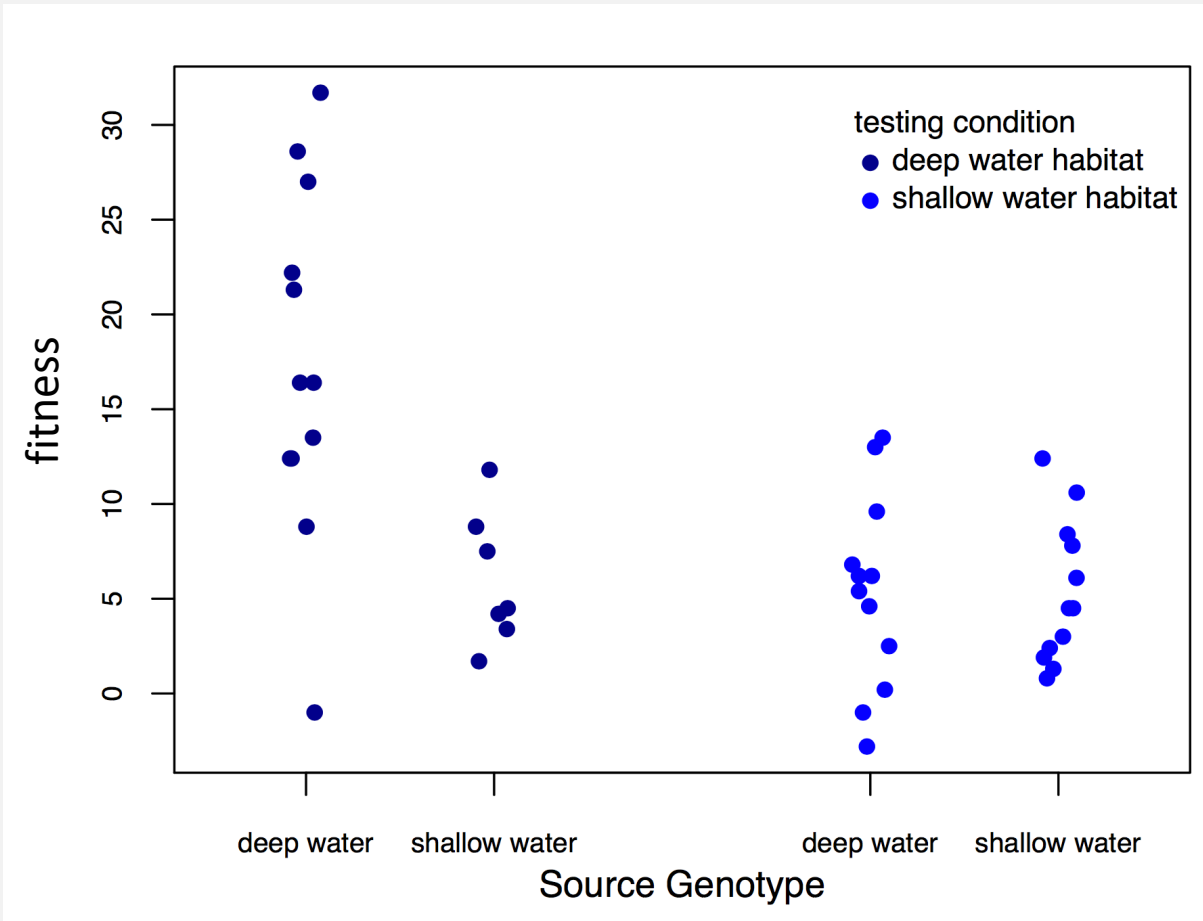
Reciprocal relocation experiment to investigate how genotype and habitat interact to determine the fitness of stickleback fish (Rundle 2002).

		Source habitat	
		Shallow	Deep
Test habitat	Shallow	12 fish	11 fish
	Deep	7 fish	11 fish





# Example of fixed effects (two factor ANOVA)



```
> anova(lm(fitness ~ genotype * test.habitat))
```

Analysis of Variance Table

Response: fitness

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	1	363.49	363.49	9.7045	0.0034403	**
test.habitat	1	550.55	550.55	14.6986	0.0004485	***
genotype:test.habitat	1	333.58	333.58	8.9059	0.0048864	**
Residuals	39	1460.77	37.46			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> summary(lm(fitness ~ genotype * test.habitat))
```

Call:

```
lm(formula = fitness ~ genotype * test.habitat)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.4750	-3.6917	-0.8083	3.4583	14.2250

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	17.475	1.767	9.891	3.49e-12	***
genotypeshallow	-11.489	2.911	-3.947	0.000321	***
test.habitatshallow	-12.125	2.499	-4.853	1.99e-05	***
genotypeshallow:test.habitatshallow	11.448	3.836	2.984	0.004886	**

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.12 on 39 degrees of freedom

Multiple R-squared: 0.4606, Adjusted R-squared: 0.4192

F-statistic: 11.1 on 3 and 39 DF, p-value: 2.093e-05

# Interpreting Coefficients

```
> summary(lm(fitness ~ genotype * test.habitat))
```

Call:

```
lm(formula = fitness ~ genotype * test.habitat)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.4750	-3.6917	-0.8083	3.4583	14.2250

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	17.475	1.767	9.891	3.49e-12	***
genotypeshallow	-11.489	2.911	-3.947	0.000321	***
test.habitatshallow	-12.125	2.499	-4.853	1.99e-05	***
genotypeshallow:test.habitatshallow	11.448	3.836	2.984	0.004886	**

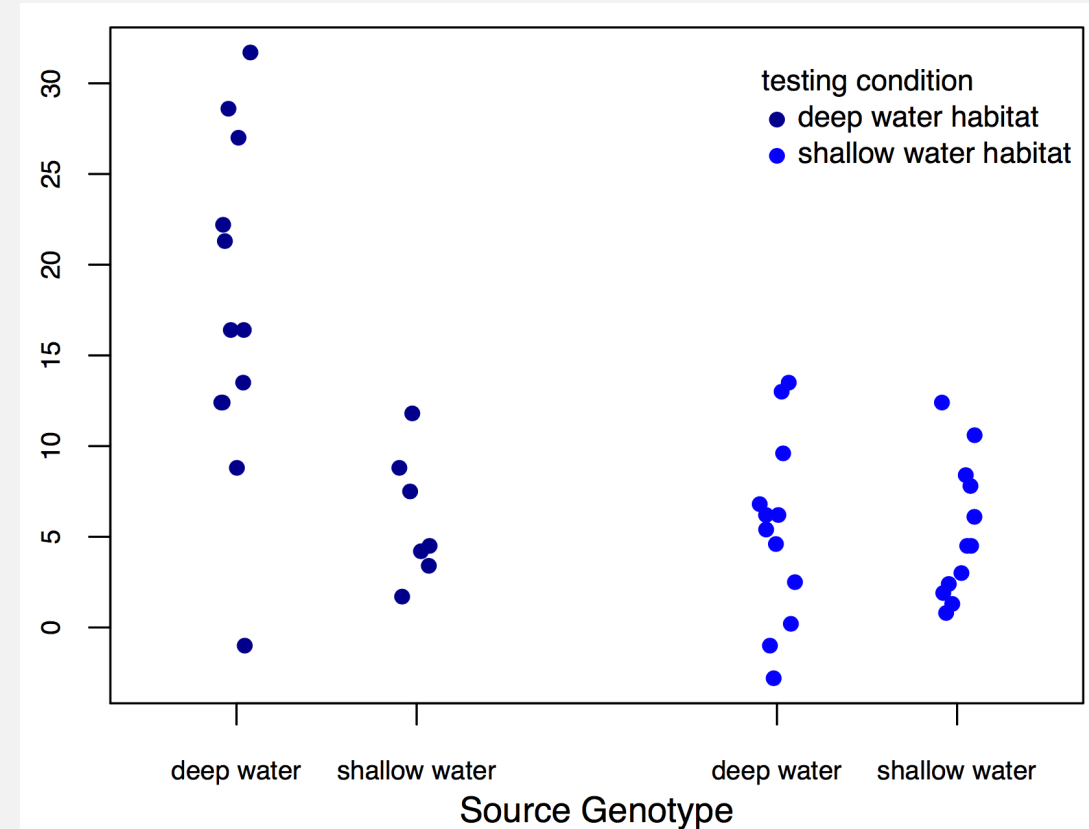
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.12 on 39 degrees of freedom

Multiple R-squared: 0.4606, Adjusted R-squared: 0.4192

F-statistic: 11.1 on 3 and 39 DF, p-value: 2.093e-05



# What is a random effect

These are randomly sampled categories of a variable that represent groups of individual measurements. Usually random effects are not repeatable.

- Study sites
- Environmental chambers
- Families made up of siblings
- Measurements within individuals

Conclusions that you reach are applicable only to the sample being studied.

# What is a random effect

Sometimes random effects are a nuisance

- Field sites
- Environmental chambers
- Field plots
- Repeated measures

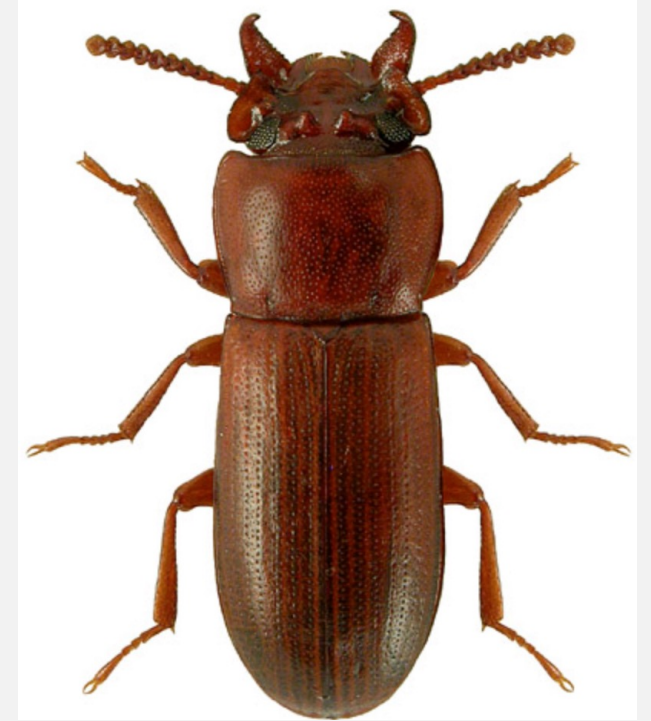

Occasionally random effects are of great interest

- Families - Heritability
- Individuals – Breeding value

# Example of random effects

Impact of selective regime on horn size.  
Measure both the left and right horn in 25  
beetles from two different selective regimes.

Horn size	Beetle	Selective regime
256	1	High
276	1	High
321	2	High
321	2	High
423	3	Low
401	3	Low
381	4	Low
409	4	Low



# Example of random effects

Identifying the predictors for the presence or absence of *Chrysina* beetles.

Number collected	oak	juniper	site	date	trip
8	1	0	21	210	A
2	1	1	13	210	A
1	0	1	31	211	A
5	0	1	15	212	A
4	1	1	21	242	B
6	1	0	13	242	B
0	1	1	31	245	B
7	1	1	15	245	B



# Implementing a mixed effects model

Mixed effect models can be fit using the LME function from the package nlme.

```
library(nlme)
fit <- lme(sqrt(beyeri) ~ oaks + jun + elev,
           random = list(~1|site,~1|trip),
           data=dat)
summary(fit)
```

Fixed effects

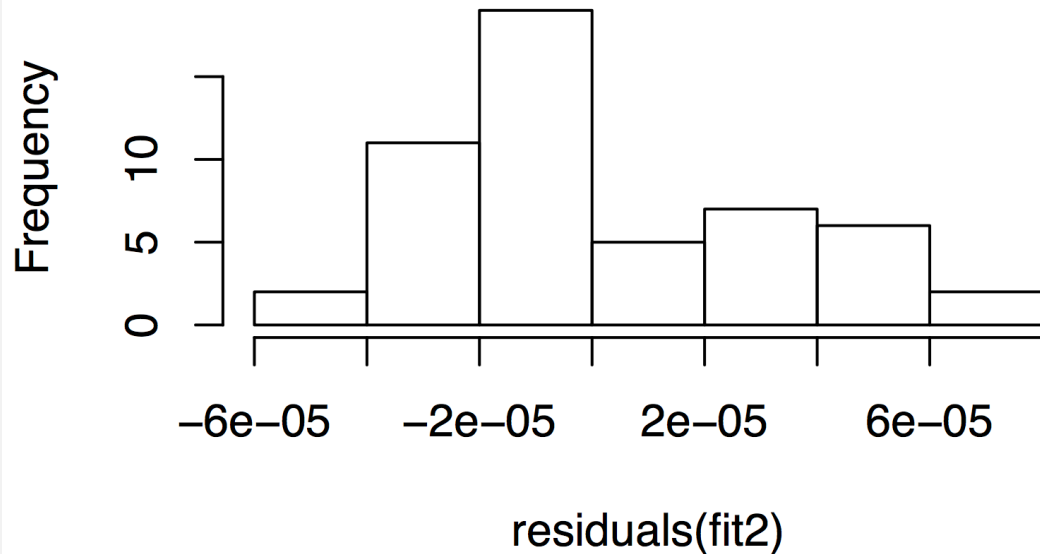
Random effects

Repeated measures at sites can't  
be treated as independent

# Implementing a mixed effects model

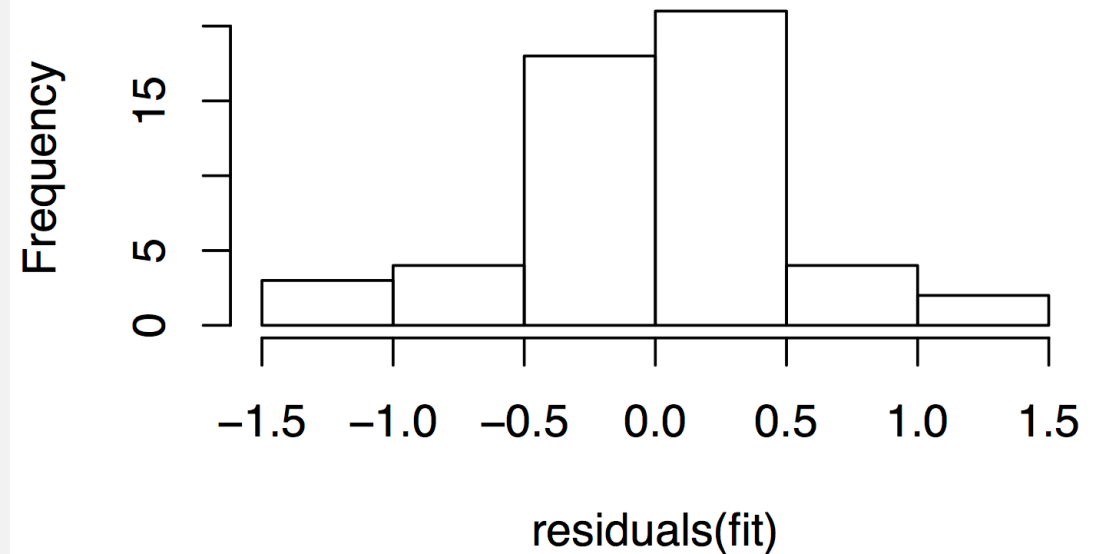
```
fit2 <- lme(beyeri ~ oaks + jun + elev,  
  random = list(~1|site, ~1|trip),  
  data=dat)
```

**Histogram of residuals(fit2)**



```
fit <- lme(sqrt(beyeri) ~ oaks + jun + elev,  
  random = list(~1|site, ~1|trip),  
  data=dat)
```

**Histogram of residuals(fit)**





# Implementing a mixed effects model

Mixed effect models can be fit using the LME function from the package nlme.

```
fit <- lme(sqrt(beyeri) ~ oaks + jun + elev,
  random = list(~1|site,~1|trip),
  data=dat)
> summary(fit)
Linear mixed-effects model fit by REML
Data: dat
      AIC      BIC    logLik
153.6247 166.7231 -69.81233

Random effects:
Formula: ~1 | site
(Intercept)
StdDev: 2.272341e-05

Formula: ~1 | trip %in% site
(Intercept) Residual
StdDev: 0.8169537 0.00264004

Fixed effects: sqrt(beyeri) ~ oaks + jun + elev
              Value Std.Error DF   t-value p-value
(Intercept) -1.6788177 1.5125428 26 -1.109931 0.2772
oaks         0.9707025 0.2943403 22  3.297892 0.0033
jun        -0.0860503 0.2460159 22 -0.349775 0.7298
elev         0.0012513 0.0008072 22  1.550167 0.1354

Correlation:
(Intr) oaks  jun
oaks  0.530
jun   0.197 -0.097
elev -0.993 -0.584 -0.250
```

```
fit <- lme(sqrt(beyeri) ~ oaks + elev,
  random = list(~1|site,~1|trip),
  data=dat)
> summary(fit)
Linear mixed-effects model fit by REML
Data: dat
      AIC      BIC    logLik
150.7722 162.1231 -69.3861

Random effects:
Formula: ~1 | site
(Intercept)
StdDev: 2.191491e-05

Formula: ~1 | trip %in% site
(Intercept) Residual
StdDev: 0.8096043 0.002598713

Fixed effects: sqrt(beyeri) ~ oaks + elev
              Value Std.Error DF   t-value p-value
(Intercept) -1.5744919 1.4695032 26 -1.071445 0.2938
oaks         0.9607576 0.2903283 23  3.309211 0.0031
elev         0.0011807 0.0007746 23  1.524404 0.1410

Correlation:
(Intr) oaks
oaks  0.563
elev -0.994 -0.631
```

```
fit <- lme(sqrt(beyeri) ~ oaks,
  random = list(~1|site,~1|trip),
  data=dat)
> summary(fit)
Linear mixed-effects model fit by REML
Data: dat
      AIC      BIC    logLik
138.5902 148.1504 -64.29512

Random effects:
Formula: ~1 | site
(Intercept)
StdDev: 2.309943e-05

Formula: ~1 | trip %in% site
(Intercept) Residual
StdDev: 0.8202517 0.002667513

Fixed effects: sqrt(beyeri) ~ oaks
              Value Std.Error DF   t-value p-value
(Intercept) 0.6514133 0.1674341 26  3.890566 6e-04
oaks        1.2400606 0.2281742 24  5.434710 0e+00

Correlation:
(Intr)
oaks -0.734
```

If you report your df with your F-statistic the reviewer will know if you did the right type of model

# Considerations for models with random effects

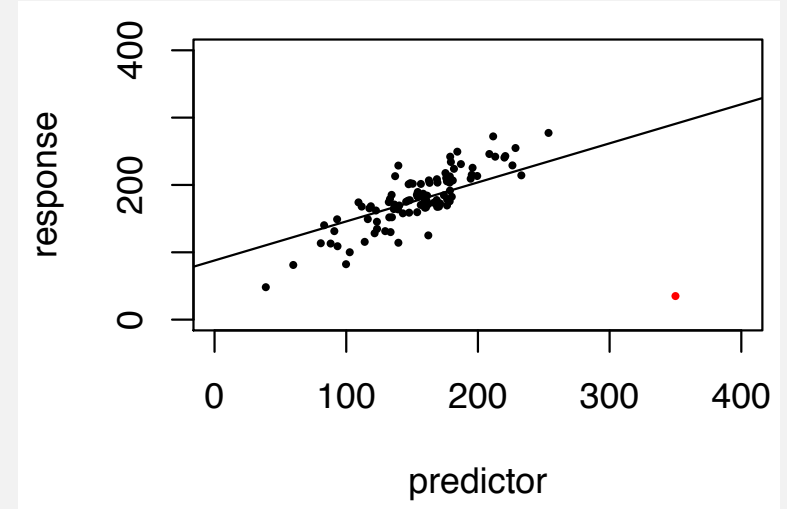
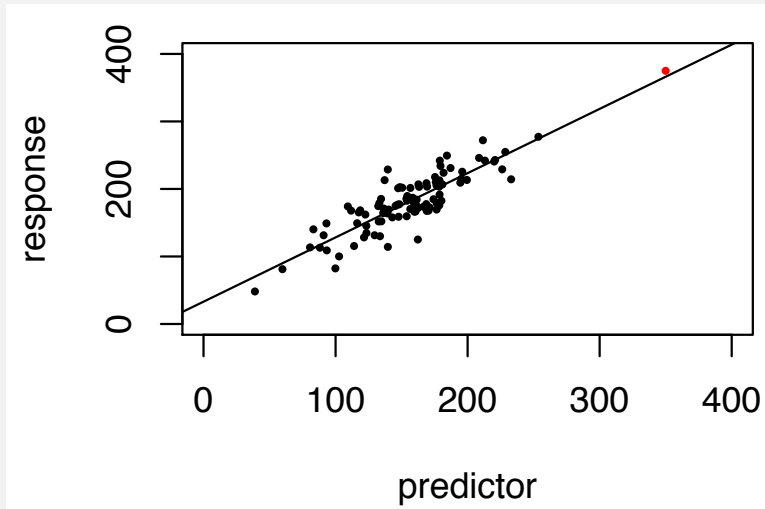
- Most software will assume that all factors are fixed unless you specify them as mixed.
- Designating factors as random effects takes extra work.
- The `lm` function treats all predictors as fixed effects.
- Treating random effects as fixed effects is fundamentally wrong.

# 4 Types of outliers

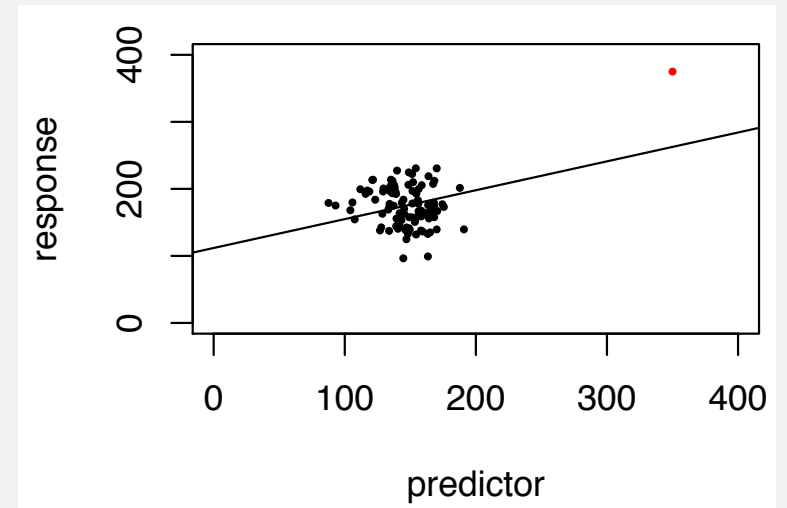
1) Obviously erroneously recorded or measured data

3) Extreme data point that impacts statistic of interest.

2) Extreme data point that doesn't impact statistic of interest but does impact p-value.



4) Extreme data point that creates significance.



# Links

[MCMCMglmm](#): Fit mixed models with phylogenetic or pedigree information in a Bayesian framework.

[Outlier Package](#): Apply outlier tests to identify possible outlier datapoints - I don't recommend this.

**Paper for next Wednesday**

**Post example code**

**No class on Monday**