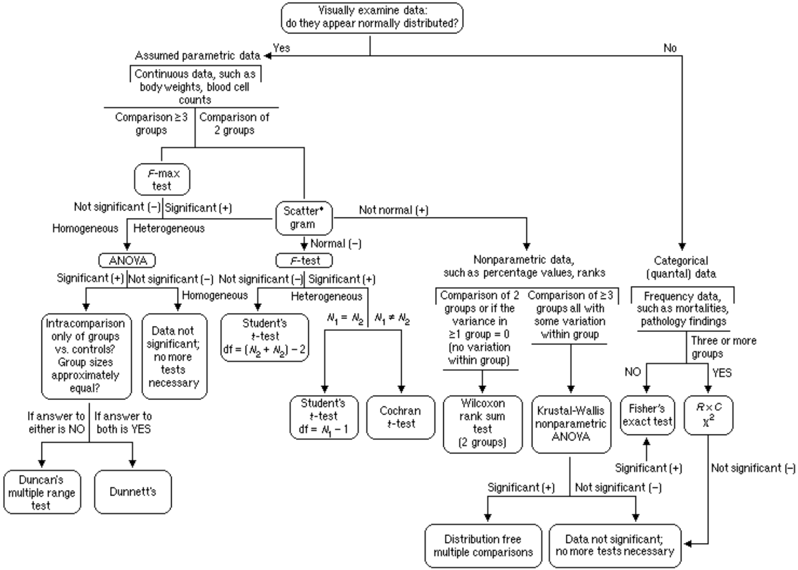
Generalized Linear Models with R

Francisco Rodriguez-Sanchez

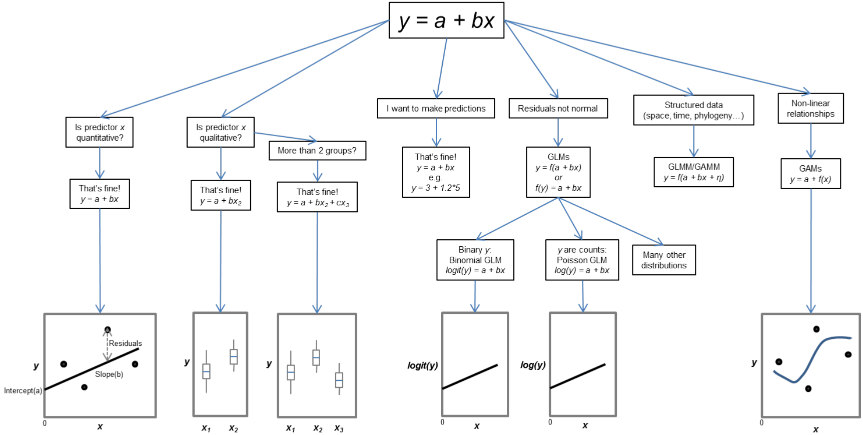
@frod_san

Introduction to linear models

Modern statistics are easier than this


Visually examine data: do they appear normally distributed?
Yes
No
Assumed parametric data
Continuous data, such as body weights, blood cell counts
Comparison ≥3 groups
Comparison of 2 groups
F-max test
Not significant (−)
Significant (+)
Homogeneous
Heterogeneous
Scatter* gram
Not normal (+)
Normal (−)
ANOVA
F-test
Significant (+)
Not significant (−)
Not significant (−)
Significant (+)
Homogeneous
Heterogeneous
Intracomparison only of groups vs. controls? Group sizes approximately equal?
Data not significant; no more tests necessary
Student's t-test df = (N2 + N2) − 2
N1 = N2
N1 ≠ N2
If answer to either is NO
If answer to both is YES
Duncan's multiple range test
Dunnett's
Student's t-test df = N1 − 1
Cochran t-test
Nonparametric data, such as percentage values, ranks
Comparison of 2 groups or if the variance in ≥1 group = 0 (no variation within group)
Comparison of ≥3 groups all with some variation within group
Wilcoxon rank sum test (2 groups)
Krustal-Wallis nonparametric ANOVA
Significant (+)
Not significant (−)
Distribution free multiple comparisons
Data not significant; no more tests necessary
Categorical (quantal) data
Frequency data, such as mortalities, pathology findings
Three or more groups
NO
YES
Fisher's exact test
R × C χ2
Significant (+)
Not significant (−)

A unified framework



$y = a + bx$

Is predictor x quantitative?

That's fine! $y = a + bx$

Is predictor x qualitative?

That's fine! $y = a + bx_2$

More than 2 groups?

That's fine! $y = a + bx_2 + cx_3$

I want to make predictions

That's fine! $y = a + bx$ e.g. $y = 3 + 1.2*5$

Residuals not normal

GLMs $y = f(a + bx)$ *or* $f(y) = a + bx$

Binary y: Binomial GLM $logit(y) = a + bx$

y are counts: Poisson GLM $log(y) = a + bx$

Many other distributions

Structured data (space, time, phylogeny…)

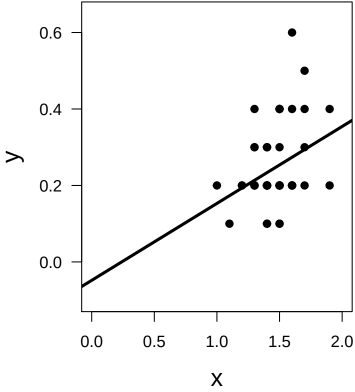
GLMM/GAMM $y = f(a + bx + \eta)$

Non-linear relationships

GAMs $y = a + f(x)$

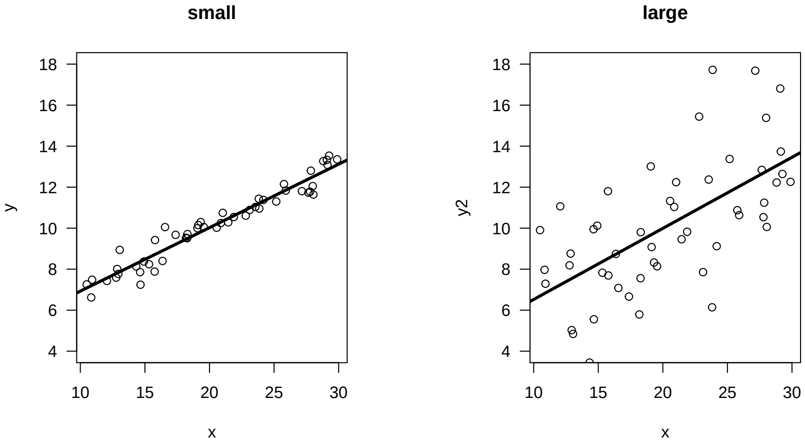
Our unified regression framework

$$y_i = a + bx_i + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$



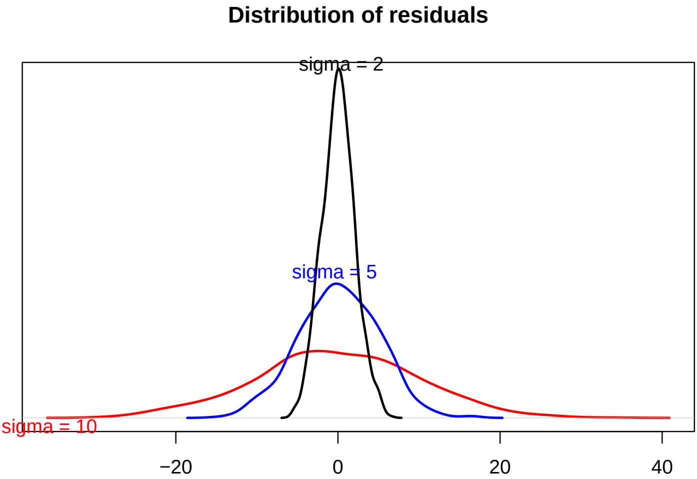
Data
y = response variable
x = predictor
Parameters
a = intercept
b = slope
σ = residual variation
ε = residuals

Residual variation (error)

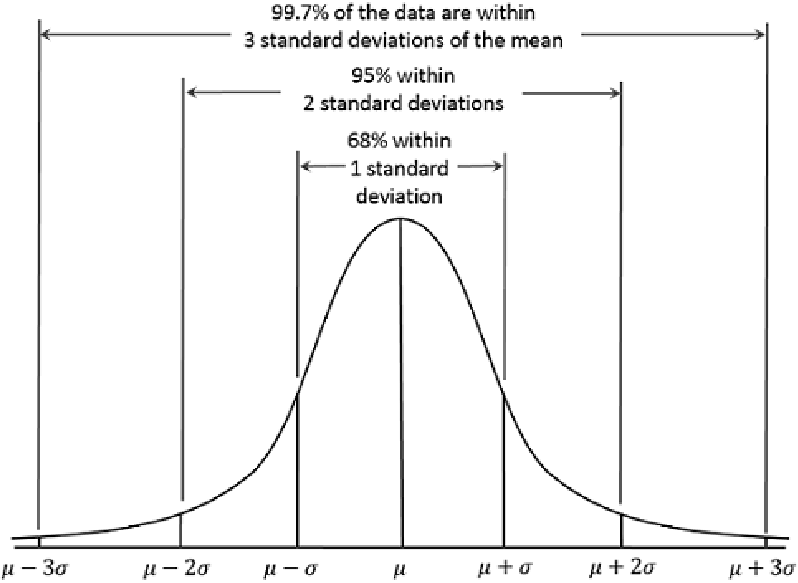


Residual variation

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$



In a Normal distribution



Different ways to write same model

$$\begin{aligned} y_i &= a + bx_i + \varepsilon_i \\ \varepsilon_i &\sim N\left(0, \sigma^2\right) \end{aligned}$$

.

$$\begin{aligned} y_i &\sim N\left(\mu_i, \sigma^2\right) \\ \mu_i &= a + bx_i \\ \varepsilon_i &\sim N\left(0, \sigma^2\right) \end{aligned}$$

Linear models

Example dataset: forest trees

- Go to https://tinyurl.com/treesdata

```r
trees <- read.csv("data-raw/trees.csv")
head(trees)
```

```
  site   dbh height    sex dead
1    4 29.68   36.1   male    0
2    5 33.29   42.3   male    0
3    2 28.03   41.9 female    0
4    5 39.86   46.5 female    0
5    1 47.94   43.9 female    0
6    1 10.82   26.2   male    0
```

Example dataset: forest trees

- Go to https://tinyurl.com/treesdata
- Download zip file and uncompress (within your project folder!)

```r
trees <- read.csv("data-raw/trees.csv")
head(trees)
```

```
  site   dbh height    sex dead
1    4 29.68   36.1   male    0
2    5 33.29   42.3   male    0
3    2 28.03   41.9 female    0
4    5 39.86   46.5 female    0
5    1 47.94   43.9 female    0
6    1 10.82   26.2   male    0
```

Questions

- What is the relationship between DBH and height?

Questions

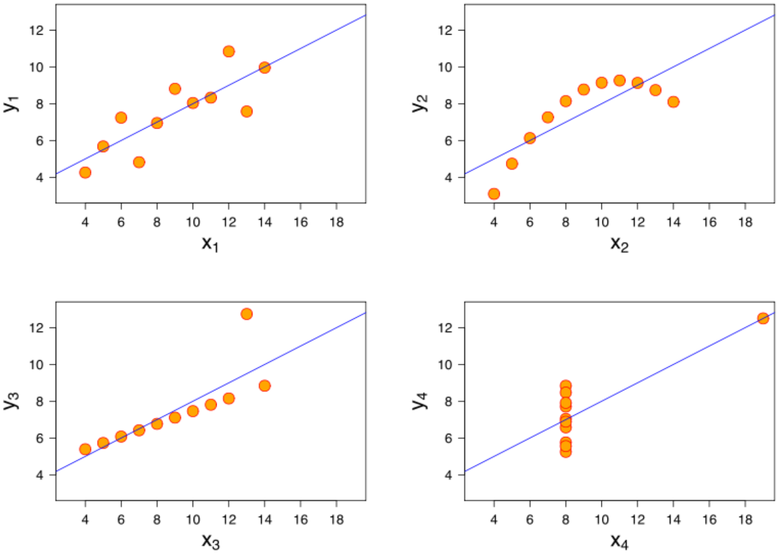
- What is the relationship between DBH and height?
- Do taller trees have bigger trunks?

Questions

- What is the relationship between DBH and height?
- Do taller trees have bigger trunks?
- Can we predict height from DBH? How well?

Always plot your data first!

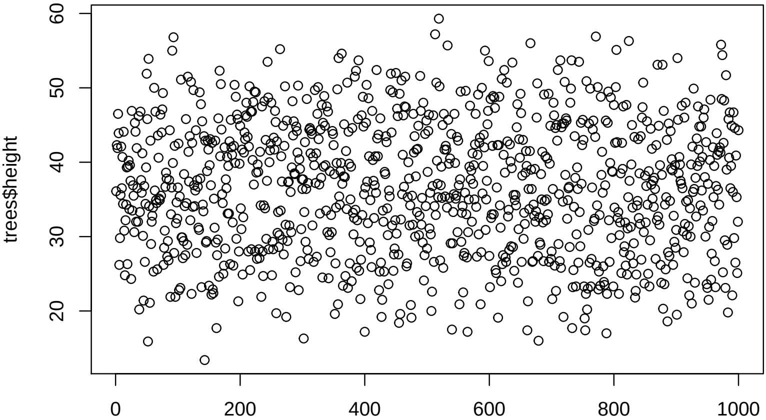
Always plot your data first!



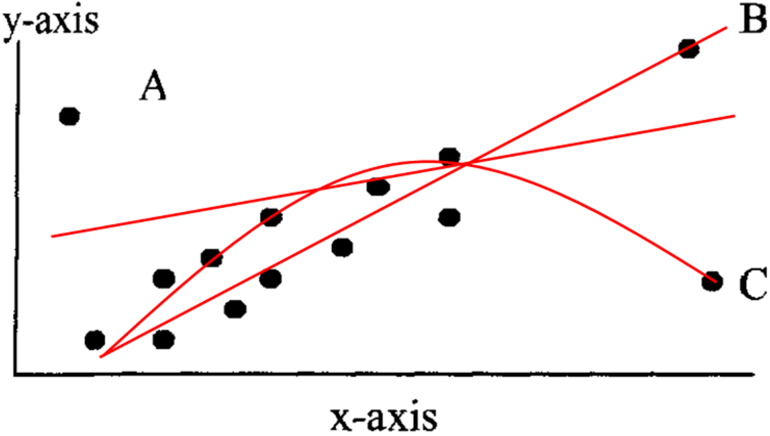
Exploratory Data Analysis (EDA)

Outliers

```
plot(trees$height)
```



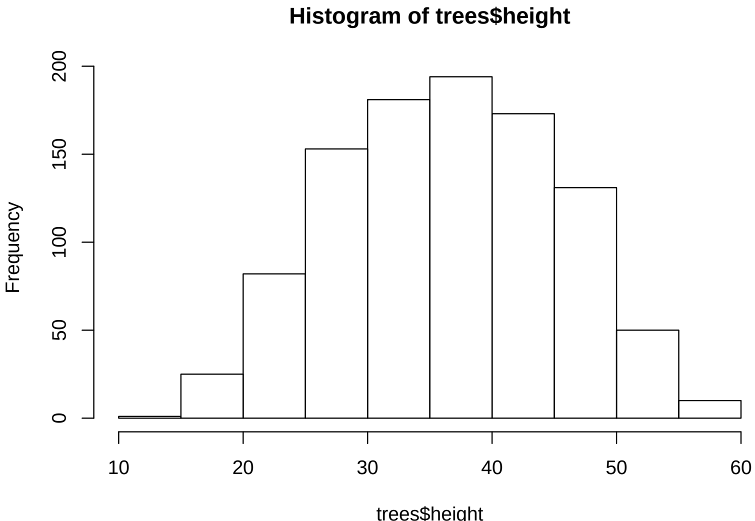
Outliers impact on regression



See http://rpsychologist.com/d3/correlation/

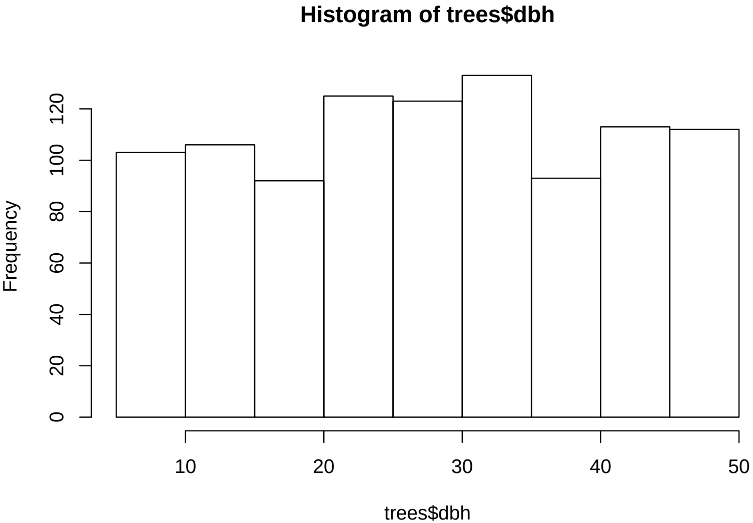
Histogram of response variable

```r
hist(trees$height)
```



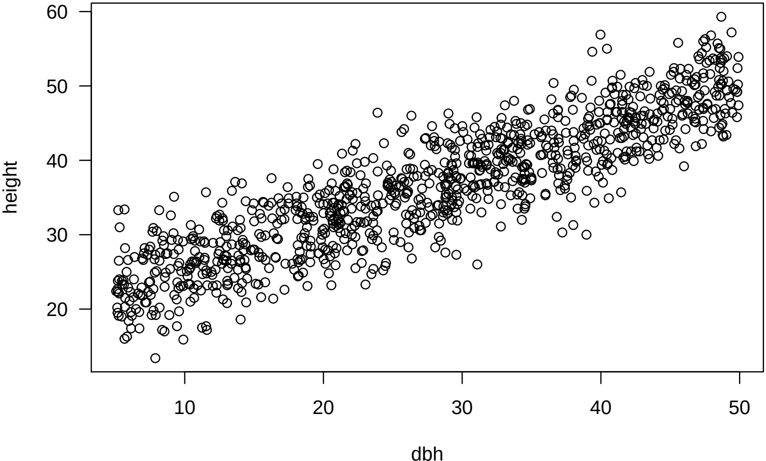
Histogram of predictor variable

```r
hist(trees$dbh)
```



Scatterplot

```
plot(height ~ dbh, data = trees, las = 1)
```



Model fitting

Now fit model

Hint: `lm`

Now fit model

Hint: `lm`

```r
m1 <- lm(height ~ dbh, data = trees)
```

which corresponds to

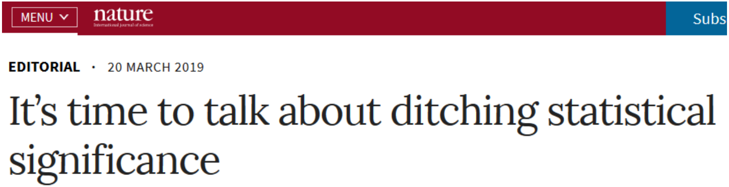
$$\begin{aligned} Height_i = a + b \cdot DBH_i + \varepsilon_i \\ \varepsilon_i \sim N\left(0, \sigma^2\right) \end{aligned}$$

Model interpretation

What does this mean?

```
Call:
lm(formula = height ~ dbh, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3270  -2.8978   0.1057   2.7924  12.9511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.33920    0.31064   62.26   <2e-16 ***
dbh          0.61570    0.01013   60.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.093 on 998 degrees of freedom
Multiple R-squared:  0.7874,    Adjusted R-squared:  0.7871
F-statistic:  3695 on 1 and 998 DF,  p-value: < 2.2e-16
```

Avoid dichotomania of statistical significance

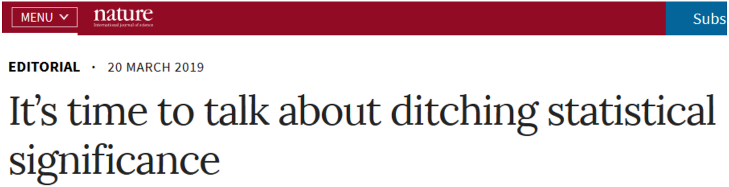


EDITORIAL · 20 MARCH 2019

It’s time to talk about ditching statistical significance

- ‘Never conclude there is ‘no difference’ or ‘no association’ just because p $>$ 0.05 or CI includes zero’

Avoid dichotomania of statistical significance

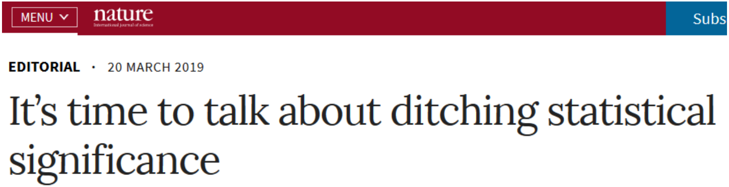


EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

- 'Never conclude there is 'no difference' or 'no association' just because $p > 0.05$ or CI includes zero'
- Estimate and communicate effect sizes and their uncertainty

Avoid dichotomania of statistical significance



EDITORIAL · 20 MARCH 2019

It’s time to talk about ditching statistical significance

- ‘Never conclude there is ‘no difference’ or ‘no association’ just because $p > 0.05$ or CI includes zero’
- Estimate and communicate effect sizes and their uncertainty
- https://doi.org/10.1038/d41586-019-00857-9

Communicating results

We found a ~~significant~~ positive relationship between DBH and Height ~~($p<0.05$)~~ ($b = 0.61$, $SE = 0.01$).

Presenting model results

```
kable(xtable::xtable(m1), digits = 2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.34	0.31	62.26	0
dbh	0.62	0.01	60.79	0

Confidence intervals

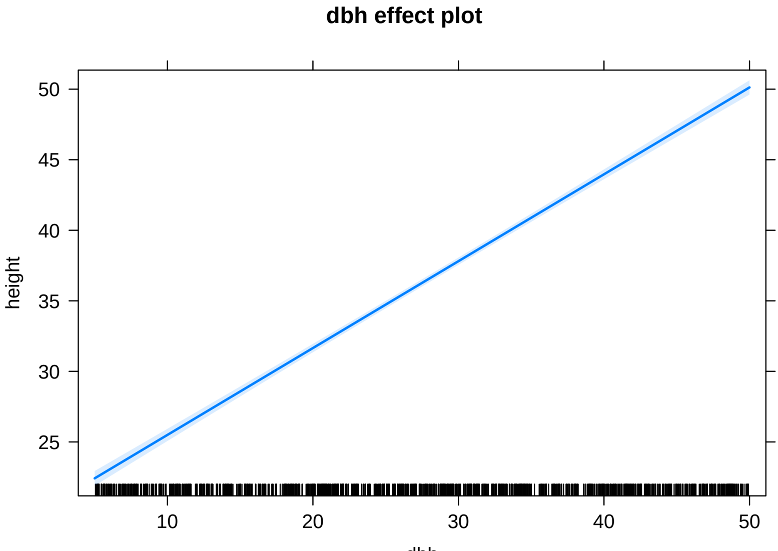
```r
confint(m1)
```

```
                2.5 %    97.5 %
(Intercept) 18.7296053 19.948788
dbh          0.5958282  0.635579
```

Visualising fitted model

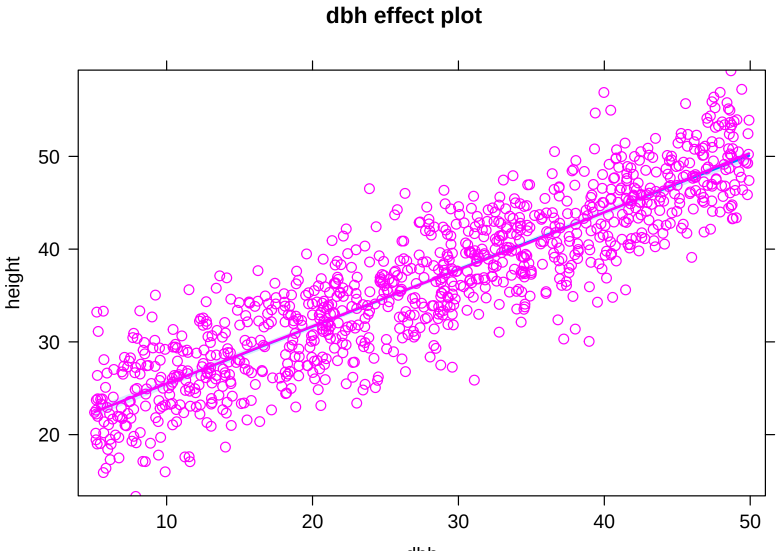
Plot effects

```
plot(allEffects(m1))
```



Plot effects

```
plot(allEffects(m1, residuals = TRUE))
```



Model checking

Linear model assumptions

- Linearity (transformations, GAM…)

Linear model assumptions

- Linearity (transformations, GAM...)
- Residuals:

Linear model assumptions

- Linearity (transformations, GAM...)
- Residuals:
  - Independent

Linear model assumptions

- Linearity (transformations, GAM...)
- Residuals:
  - Independent
  - Equal variance

Linear model assumptions

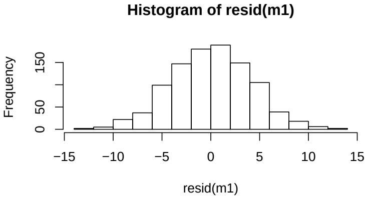
- Linearity (transformations, GAM…)
- Residuals:
  - Independent
  - Equal variance
  - Normal

Linear model assumptions

- Linearity (transformations, GAM…)
- Residuals:
  - Independent
  - Equal variance
  - Normal
- No measurement error in predictors

Are residuals normal?

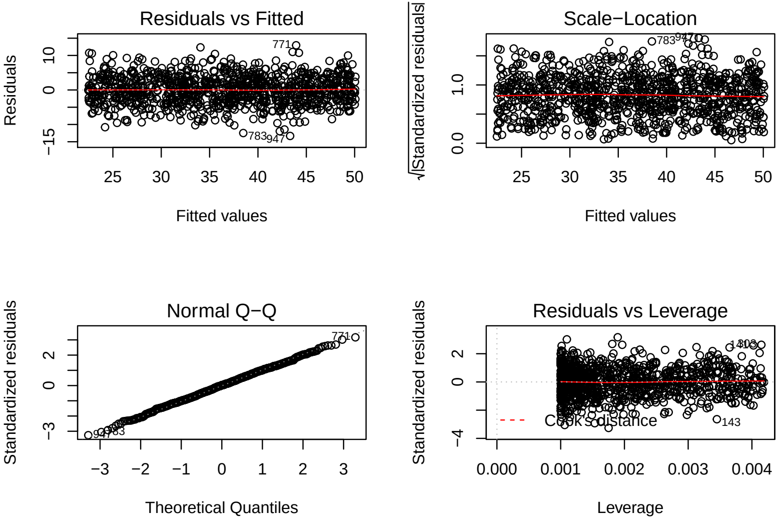
```
hist(resid(m1))
```



```
lm(formula = height ~ dbh, data = trees)
            coef.est coef.se
(Intercept) 19.34     0.31
dbh          0.62     0.01
---
n = 1000, k = 2
residual sd = 4.09, R-Squared = 0.79
```

SD of residuals = 4.09 coincides with estimate of `sigma`.

Model checking: residuals



Using model for prediction

How good is the model in predicting tree height?

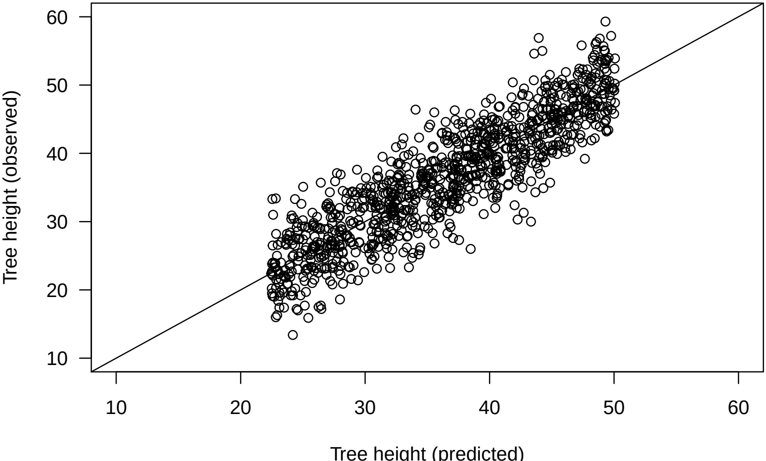
`fitted` gives predictions for each observation

```
trees$height.pred <- fitted(m1)
head(trees)
```

```
  site   dbh height    sex dead height.pred
1    4 29.68   36.1   male    0    37.61328
2    5 33.29   42.3   male    0    39.83597
3    2 28.03   41.9 female    0    36.59737
4    5 39.86   46.5 female    0    43.88114
5    1 47.94   43.9 female    0    48.85603
6    1 10.82   26.2   male    0    26.00111
```

Calibration plot: Observed vs Predicted values

```
plot(trees$height.pred, trees$height, xlab = "Tree height (predi
```



Workflow

- **Visualise data**

Workflow

- **Visualise data**
- **Understand fitted model** (`summary`, `allEffects`…)

Workflow

- **Visualise data**
- **Understand fitted model** (`summary`, `allEffects`…)
- **Visualise model** (`plot(allEffects)`, `visreg`, `plot_model`…)

Workflow

- **Visualise data**
- **Understand fitted model** (`summary`, `allEffects`…)
- **Visualise model** (`plot(allEffects)`, `visreg`, `plot_model`…)
- **Check model** (`plot`, `resid_panel`, calibration plot…)

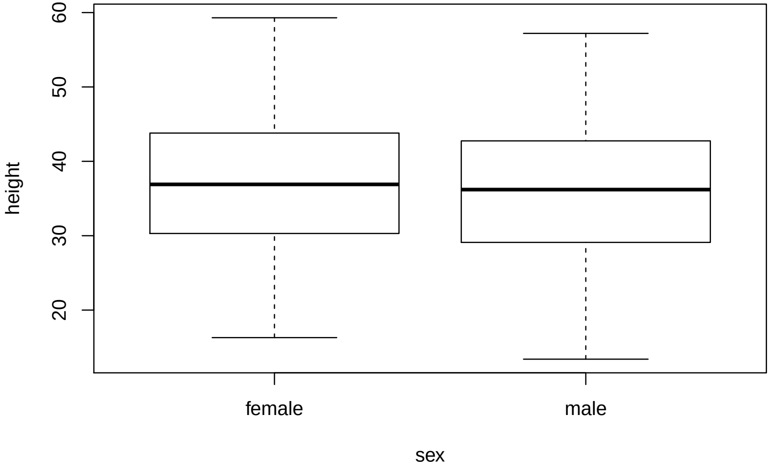
Workflow

- **Visualise data**
- **Understand fitted model** (`summary`, `allEffects`...)
- **Visualise model** (`plot(allEffects)`, `visreg`, `plot_model`...)
- **Check model** (`plot`, `resid_panel`, calibration plot...)
- **Predict** (`fitted`, `predict`)

Categorical predictors (factors)

Q: Does tree height vary with sex?

```r
plot(height ~ sex, data = trees)
```



Model height ~ sex

```
m2 <- lm(height ~ sex, data = trees)
```

```

Call:
lm(formula = height ~ sex, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.6881  -6.7881  -0.0097   6.7261  22.3687 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.9312     0.3981  92.778   <2e-16 ***
sexmale      -0.8432     0.5607  -1.504    0.133    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.865 on 998 degrees of freedom
Multiple R-squared:  0.002261,  Adjusted R-squared:  0.001261 
F-statistic: 2.261 on 1 and 998 DF,  p-value: 0.133
```

Linear model with categorical predictors

```r
m2 <- lm(height ~ sex, data = trees)
```

corresponds to

$$
\begin{aligned}
Height_i = a + b_{male} + \varepsilon_i \\
\varepsilon_i \sim N\left(0, \sigma^2\right)
\end{aligned}
$$

Model height ~ sex

```r
m2 <- lm(height ~ sex, data = trees)
```

```
Call:
lm(formula = height ~ sex, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.6881  -6.7881  -0.0097   6.7261  22.3687 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.9312     0.3981  92.778   <2e-16 ***
sexmale      -0.8432     0.5607  -1.504    0.133    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.865 on 998 degrees of freedom
Multiple R-squared:  0.002261,  Adjusted R-squared:  0.001261 
F-statistic: 2.261 on 1 and 998 DF,  p-value: 0.133
```

Presenting model results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.93	0.40	92.78	0.00
sexmale	-0.84	0.56	-1.50	0.13

Effects: Height $\sim$ sex

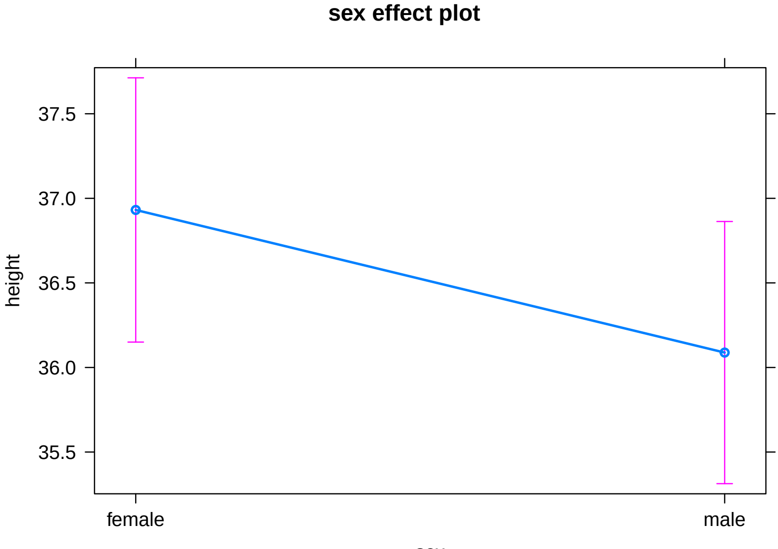
Compare CIs

```r
summary(allEffects(m2))
```

```
 model: height ~ sex

 sex effect
sex
  female     male 
36.93125 36.08810 

 Lower 95 Percent Confidence Limits
sex
  female     male 
36.15012 35.31319 

 Upper 95 Percent Confidence Limits
sex
  female     male 
37.71238 36.86300 
```

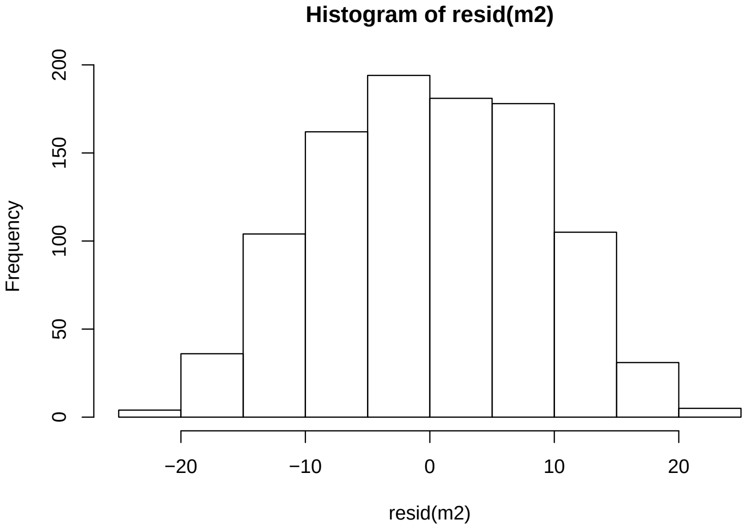
Plot

```
plot(allEffects(m2))
```



Model checking: residuals

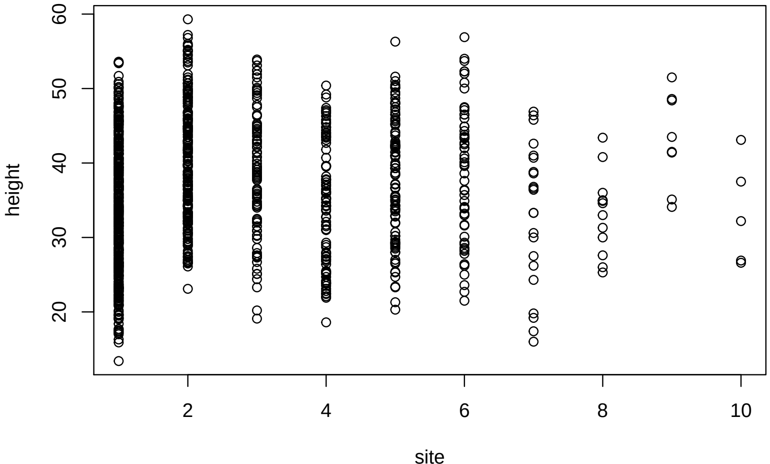
```r
hist(resid(m2))
```



Q: Does height differ among field sites?

Plot data first

```
plot(height ~ site, data = trees)
```



Linear model with categorical predictors

```
m3 <- lm(height ~ site, data = trees)
```

$$
\begin{aligned}
y_i = a + b_{site2} + c_{site3} + d_{site4} + e_{site5} + ... + \varepsilon_i \\
\varepsilon_i \sim N\left(0, \sigma^2\right)
\end{aligned}
$$

Model Height ~ site

All right here?

```r
m3 <- lm(height ~ site, data = trees)
```

```

Call:
lm(formula = height ~ site, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.4498  -6.7049   0.0709   6.7537  23.0640 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  35.4636     0.4730  74.975  < 2e-16 ***
site          0.3862     0.1413   2.733  0.00639 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.842 on 998 degrees of freedom
Multiple R-squared:  0.007429,	Adjusted R-squared:  0.006435 
```

site is a factor!

```r
trees$site <- as.factor(trees$site)
```

Model Height $\sim$ site

```
Call:
lm(formula = height ~ site, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.4416  -6.9004   0.0379   6.3051  19.7584 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  33.8416     0.4266  79.329  < 2e-16 ***
site2         6.3411     0.7126   8.899  < 2e-16 ***
site3         4.9991     0.9828   5.086 4.36e-07 ***
site4         0.5329     0.9872   0.540  0.58949    
site5         4.3723     0.9425   4.639 3.97e-06 ***
site6         4.7601     1.1709   4.065 5.18e-05 ***
site7        -0.7416     1.8506  -0.401  0.68871    
site8        -0.6832     2.4753  -0.276  0.78258    
site9         9.1709     3.0165   3.040  0.00243 ** 
site10       -0.5816     3.8013  -0.153  0.87843    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 8.446 on 990 degrees of freedom
Multiple R-squared:  0.1016,    Adjusted R-squared:  0.09344 
F-statistic: 12.44 on 9 and 990 DF,  p-value: < 2.2e-16
```

Presenting model results

```r
kable(xtable::xtable(m3), digits = 2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.84	0.43	79.33	0.00
site2	6.34	0.71	8.90	0.00
site3	5.00	0.98	5.09	0.00
site4	0.53	0.99	0.54	0.59
site5	4.37	0.94	4.64	0.00
site6	4.76	1.17	4.07	0.00
site7	-0.74	1.85	-0.40	0.69
site8	-0.68	2.48	-0.28	0.78
site9	9.17	3.02	3.04	0.00
site10	-0.58	3.80	-0.15	0.88

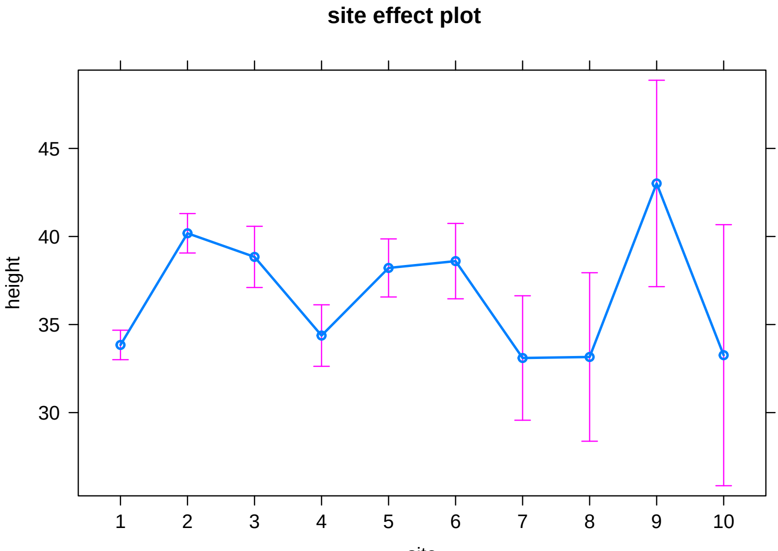
Estimated tree heights for each site

```r
summary(allEffects(m3))
```

```
 model: height ~ site

 site effect
site
       1        2        3        4        5        6        7        8 
33.84158 40.18265 38.84066 34.37444 38.21386 38.60167 33.10000 33.15833 
       9       10 
43.01250 33.26000 

 Lower 95 Percent Confidence Limits
site
       1        2        3        4        5        6        7        8 
33.00444 39.06264 37.10317 32.62733 36.56463 36.46190 29.56629 28.37367 
       9       10 
37.15251 25.84764 

 Upper 95 Percent Confidence Limits
site
       1        2        3        4        5        6        7        8 
34.67872 41.30265 40.57814 36.12156 39.86309 40.74143 36.63371 37.94299 
       9       10 
48.87249 40.67236 
```

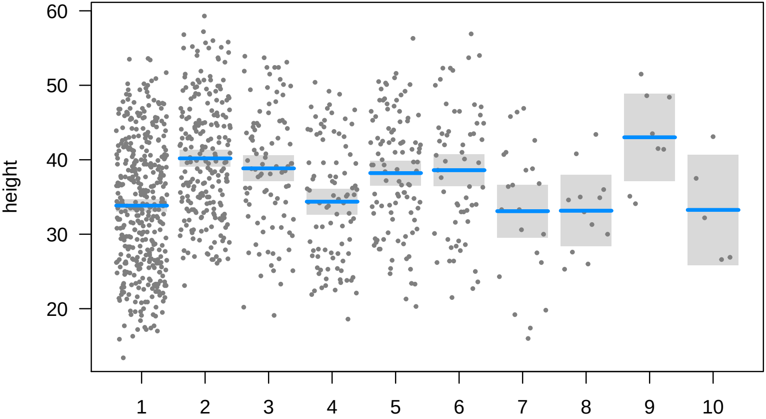
Plot

```r
plot(allEffects(m3))
```

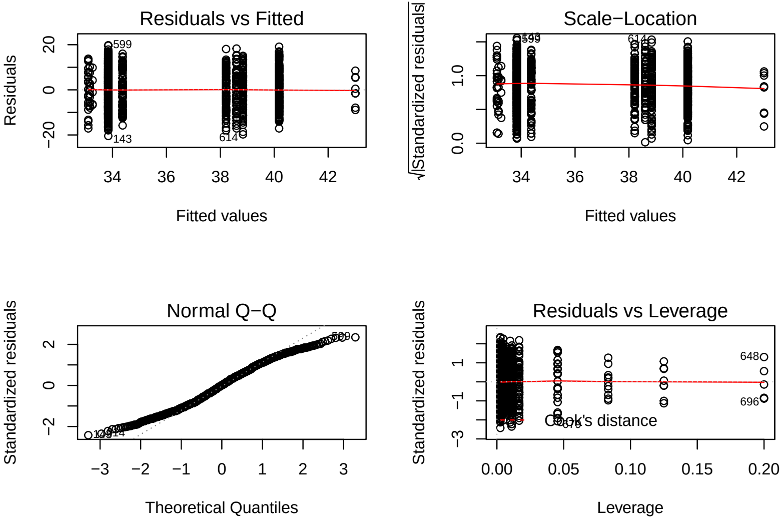


Plot (visreg)

```r
library(visreg)
visreg(m3)
```



Model checking: residuals



Combining continuous and categorical predictors

Predicting tree height based on dbh and site

```r
lm(height ~ site + dbh, data = trees)
```

corresponds to

$$y_i = a + b_{site2} + c_{site3} + d_{site4} + e_{site5} + ... + k \cdot DBH_i + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

Predicting tree height based on dbh and site

```
Call:
lm(formula = height ~ site + dbh, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1130  -1.9885   0.0582   2.0314  11.3320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.699037   0.260565  64.088  < 2e-16 ***
site2        6.504303   0.256730  25.335  < 2e-16 ***
site3        4.357457   0.354181  12.303  < 2e-16 ***
site4        1.934650   0.356102   5.433 6.98e-08 ***
site5        3.637432   0.339688  10.708  < 2e-16 ***
site6        4.204511   0.421906   9.966  < 2e-16 ***
site7       -0.176193   0.666772  -0.264   0.7916
site8       -5.312648   0.893603  -5.945 3.82e-09 ***
site9        5.437049   1.087766   4.998 6.84e-07 ***
site10       2.263338   1.369986   1.652   0.0988 .
dbh          0.617075   0.007574  81.473  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.043 on 989 degrees of freedom
Multiple R-squared:  0.8835,    Adjusted R-squared:  0.8823
F-statistic:   750 on 10 and 989 DF,  p-value: < 2.2e-16
```

Presenting model results

```r
kable(xtable::xtable(m4), digits = 2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.70	0.26	64.09	0.00
site2	6.50	0.26	25.34	0.00
site3	4.36	0.35	12.30	0.00
site4	1.93	0.36	5.43	0.00
site5	3.64	0.34	10.71	0.00
site6	4.20	0.42	9.97	0.00
site7	-0.18	0.67	-0.26	0.79
site8	-5.31	0.89	-5.95	0.00
site9	5.44	1.09	5.00	0.00
site10	2.26	1.37	1.65	0.10
dbh	0.62	0.01	81.47	0.00

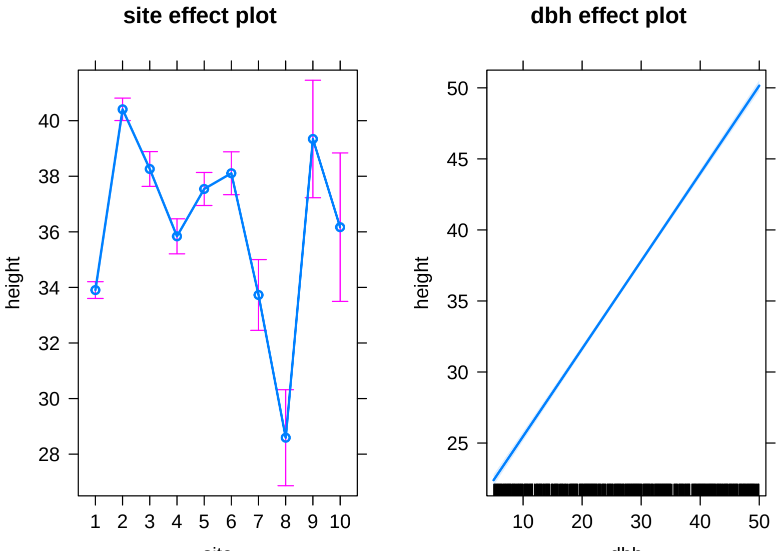
Estimated tree heights for each site

```
summary(allEffects(m4))
```

```
 model: height ~ site + dbh

 site effect
site
       1        2        3        4        5        6        7        8 
33.90437 40.40868 38.26183 35.83902 37.54181 38.10889 33.72818 28.59173 
       9       10 
39.34142 36.16771 

 Lower 95 Percent Confidence Limits
site
       1        2        3        4        5        6        7        8 
33.60276 40.00512 37.63569 35.20858 36.94739 37.33787 32.45495 26.86438 
       9       10 
37.22831 33.49623 

 Upper 95 Percent Confidence Limits
site
       1        2        3        4        5        6        7        8 
34.20599 40.81223 38.88798 36.46947 38.13622 38.87990 35.00141 30.31907 
       9       10 
41.45454 38.83919 

 dbh effect
dbh
       5       20       30       40       50 
22.38634 31.64246 37.81321 43.98396 50.15471 

 Lower 95 Percent Confidence Limits
```

Plot

```
plot(allEffects(m4))
```

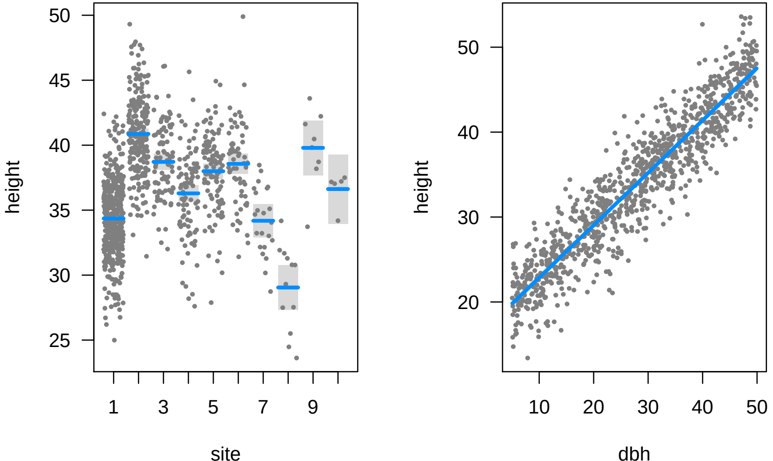


site effect plot

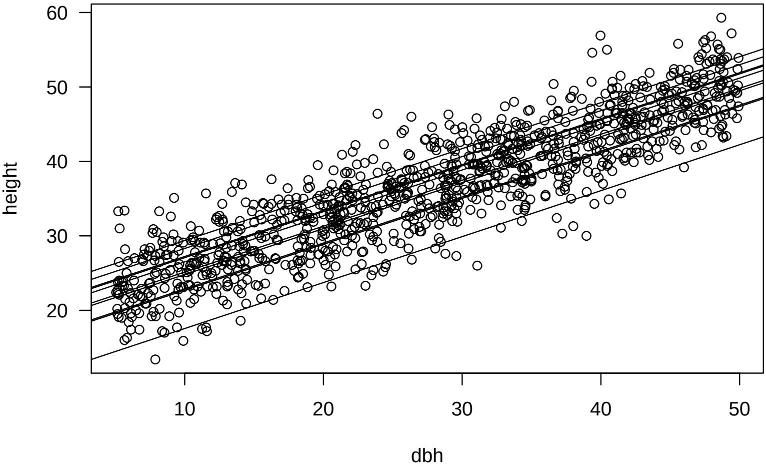
dbh effect plot

Plot (visreg)

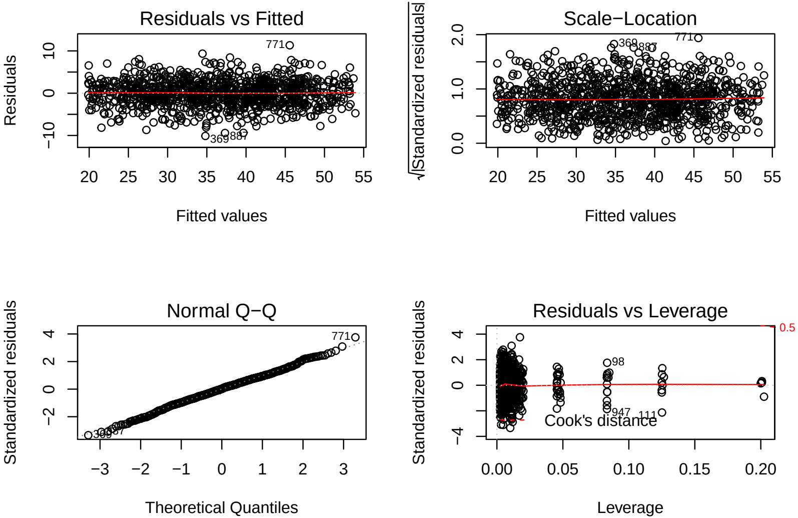
```r
visreg(m4)
```



We have fitted model w/ many intercepts and single slope

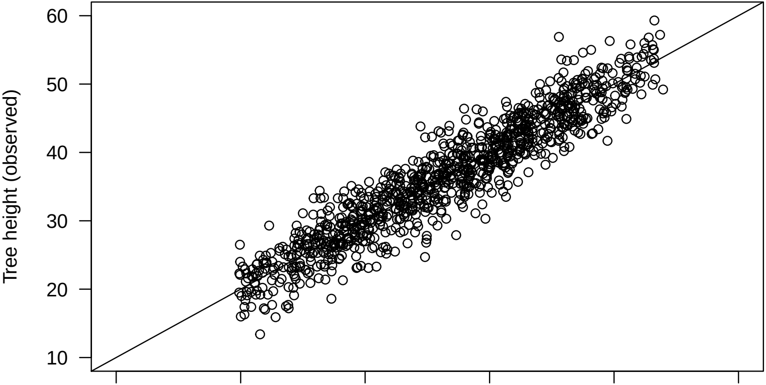


Model checking: residuals



How good is this model? Calibration plot

```
trees$height.pred <- fitted(m4)
plot(trees$height.pred, trees$height, xlab = "Tree height (predi
abline(a = 0, b = 1)
```



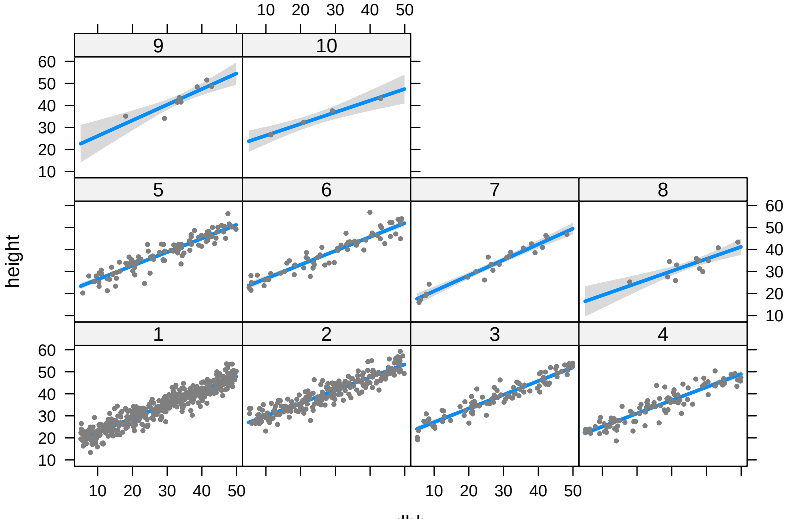
Q: Does allometric relationship between DBH and Height vary among sites?

Model with interactions

```
Call:
lm(formula = height ~ site * dbh, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1017  -1.9839   0.0645   2.0486  11.1789 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.359437   0.360054  45.436  < 2e-16 ***
site2        7.684781   0.609657  12.605  < 2e-16 ***
site3        4.518568   0.867008   5.212 2.28e-07 ***
site4        2.769336   0.813259   3.405 0.000688 ***
site5        3.917607   0.870983   4.498 7.68e-06 ***
site6        4.155161   1.009379   4.117 4.17e-05 ***
site7       -2.306799   1.551303  -1.487 0.137334    
site8       -2.616095   4.090671  -0.640 0.522630    
site9        2.621560   5.073794   0.517 0.605492    
site10       4.662340   2.991072   1.559 0.119378    
dbh          0.629299   0.011722  53.685  < 2e-16 ***
site2:dbh   -0.042784   0.020033  -2.136 0.032950 *  
site3:dbh   -0.006031   0.027640  -0.218 0.827312    
site4:dbh   -0.031633   0.028225  -1.121 0.262677    
site5:dbh   -0.010173   0.027887  -0.365 0.715334    
site6:dbh    0.001337   0.032109   0.042 0.966797    
site7:dbh    0.079728   0.052056   1.532 0.125951    
site8:dbh   -0.079027   0.113386  -0.697 0.485984    
site9:dbh    0.081035   0.146649   0.553 0.580679    
site10:dbh  -0.101107   0.114520  -0.883 0.377522    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.041 on 980 degrees of freedom
Multiple R-squared:  0.8847,    Adjusted R-squared:  0.8825 
F-statistic: 395.7 on 19 and 980 DF,  p-value: < 2.2e-16
```

Does slope vary among forests?

```
visreg(m5, xvar = "dbh", by = "site")
```



Extra exercises

- paperplanes: How does flight distance differ with age, gender or paper type?

Extra exercises

- paperplanes: How does flight distance differ with age, gender or paper type?
- mammal sleep: Are sleep patterns related to diet?

Extra exercises

- paperplanes: How does flight distance differ with age, gender or paper type?
- mammal sleep: Are sleep patterns related to diet?
- iris: Predict petal length $\sim$ petal width and species

Extra exercises

- paperplanes: How does flight distance differ with age, gender or paper type?
- mammal sleep: Are sleep patterns related to diet?
- iris: Predict petal length $\sim$ petal width and species
- racing pigeons: is speed related to sex?

Generalised Linear Models: Logistic regression

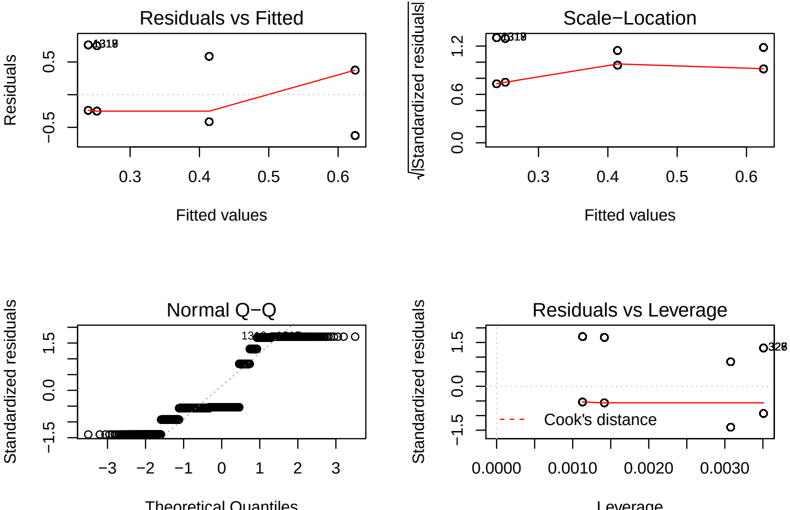
Q: Survival of passengers on the Titanic $\sim$ Class

Read `titanic_long.csv` dataset.

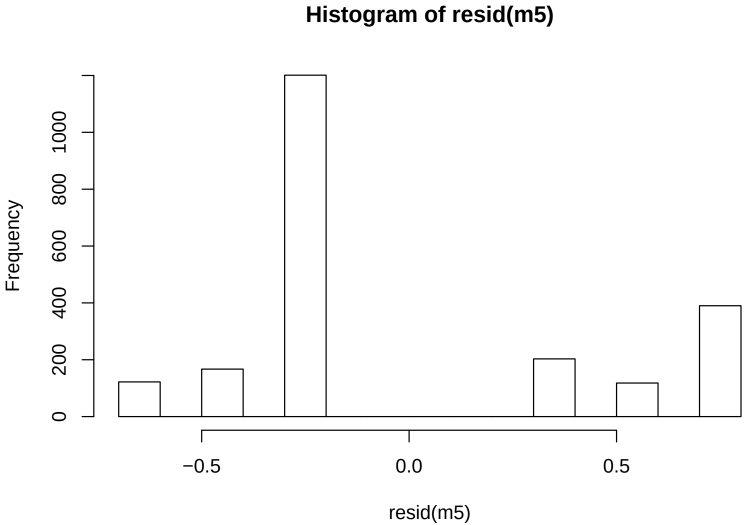
```
  class   age  sex survived
1 first adult male        1
2 first adult male        1
3 first adult male        1
4 first adult male        1
5 first adult male        1
6 first adult male        1
```

Let's fit linear model:

```r
m5 <- lm(survived ~ class, data = titanic)
```



Weird residuals!



What if your residuals are clearly non-normal, or variance not constant (heteroscedasticity)?

- Binary variables (0/1)

What if your residuals are clearly non-normal, or variance not constant (heteroscedasticity)?

- Binary variables (0/1)
- Counts (0, 1, 2, 3, ...)

Generalised Linear Models

1. **Response variable** - distribution `family`

Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial

Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial
   - Poisson

Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial
   - Poisson
   - Gamma

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc
2. **Predictors** (continuous or categorical)

Generalised Linear Models

1. **Response variable** - distribution `family`
    - Bernouilli - Binomial
    - Poisson
    - Gamma
    - etc
2. **Predictors** (continuous or categorical)
3. **Link function**

Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial
   - Poisson
   - Gamma
   - etc
2. **Predictors** (continuous or categorical)
3. **Link function**
   - Gaussian: identity

Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial
   - Poisson
   - Gamma
   - etc
2. **Predictors** (continuous or categorical)
3. **Link function**
   - Gaussian: identity
   - Binomial: logit, probit

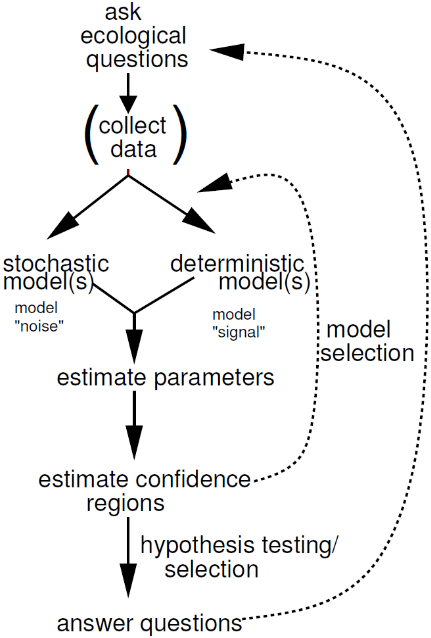
Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial
   - Poisson
   - Gamma
   - etc
2. **Predictors** (continuous or categorical)
3. **Link function**
   - Gaussian: identity
   - Binomial: logit, probit
   - Poisson: log...

Generalised Linear Models

1. **Response variable** - distribution `family`
   - Bernouilli - Binomial
   - Poisson
   - Gamma
   - etc
2. **Predictors** (continuous or categorical)
3. **Link function**
   - Gaussian: identity
   - Binomial: logit, probit
   - Poisson: log...
   - See `family`.

The modelling process



Bernouilli - Binomial distribution (Logistic regression)

- Response variable: Yes/No (e.g. survival, sex, presence/absence)

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

Then

$$\begin{aligned} Pr(alive) &= a + bx \\ logit(Pr(alive)) &= a + bx \\ Pr(alive) = invlogit(a + bx) &= \frac{e^{a+bx}}{1 + e^{a+bx}} \end{aligned}$$

Bernouilli - Binomial distribution (Logistic regression)

- Response variable: Yes/No (e.g. survival, sex, presence/absence)
- Link function: `logit` (others possible, see `family`).

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

Then

$$\begin{aligned} Pr(alive) &= a + bx \\ logit(Pr(alive)) &= a + bx \\ Pr(alive) = invlogit(a + bx) &= \frac{e^{a+bx}}{1 + e^{a+bx}} \end{aligned}$$

Back to survival of Titanic passengers

How many survived in each class?

```
table(titanic$class, titanic$survived)
```

```
         0   1
  crew   673 212
  first  122 203
  second 167 118
  third  528 178
```

Back to survival of Titanic passengers (dplyr)

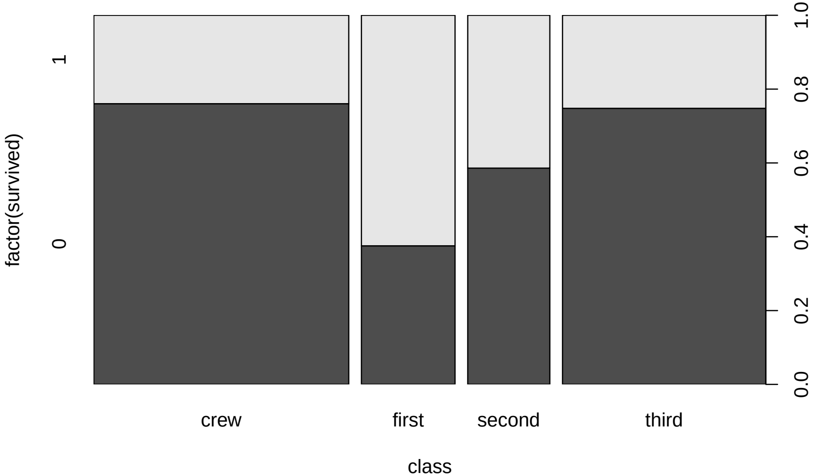
Passenger survival according to class

```
titanic %>%
  group_by(class, survived) %>%
  summarise(count = n())
```

```
# A tibble: 8 x 3
# Groups:   class [4]
  class  survived count
  <fct>     <int> <int>
1 crew          0   673
2 crew          1   212
3 first         0   122
4 first         1   203
5 second        0   167
6 second        1   118
7 third         0   528
8 third         1   178
```

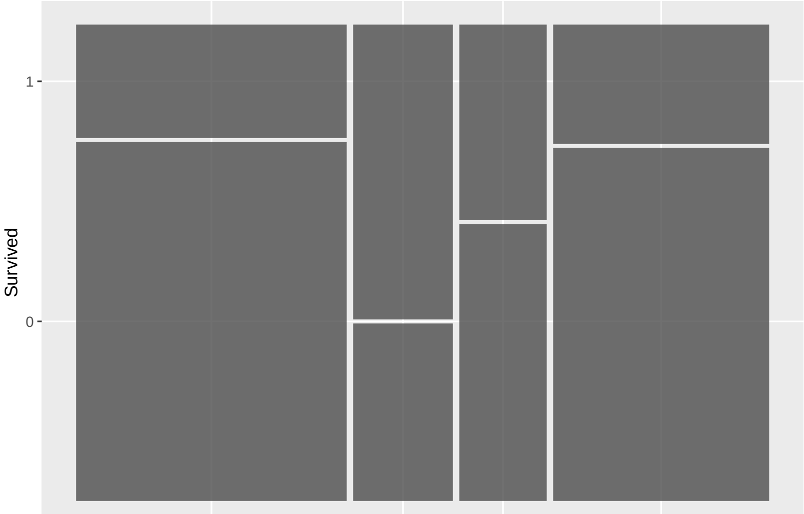
Or graphically...

```r
plot(factor(survived) ~ class, data = titanic)
```



Mosaic plots (ggplot2)

```
ggplot(titanic) +
  geom_mosaic(aes(x = product(survived, class))) +
  labs(x = "", y = "Survived")
```



Fitting GLMs in R: `glm`

```r
tit.glm <- glm(survived ~ class, data = titanic, family = binomial)
```

which corresponds to

$$logit(Pr(survival)_i) = a + b \cdot class_i$$

$$logit(Pr(survival)_i) = a + b_{first} + c_{second} + d_{third}$$

Fitting GLMs in R: glm

```r
tit.glm <- glm(survived ~ class, data = titanic, family = binomial)
```

```

Call:
glm(formula = survived ~ class, family = binomial, data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3999  -0.7623  -0.7401   0.9702   1.6906

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.15516    0.07876 -14.667  < 2e-16 ***
classfirst   1.66434    0.13902  11.972  < 2e-16 ***
classsecond  0.80785    0.14375   5.620 1.91e-08 ***
classthird   0.06785    0.11711   0.579    0.562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2588.6  on 2197  degrees of freedom
AIC: 2596.6

Number of Fisher Scoring iterations: 4
```

These estimates are in logit scale!

Model interpretation using effects package

```r
library(effects)
allEffects(tit.glm)
```

```
 model: survived ~ class

 class effect
class
     crew     first    second     third 
0.2395480 0.6246154 0.4140351 0.2521246 
```

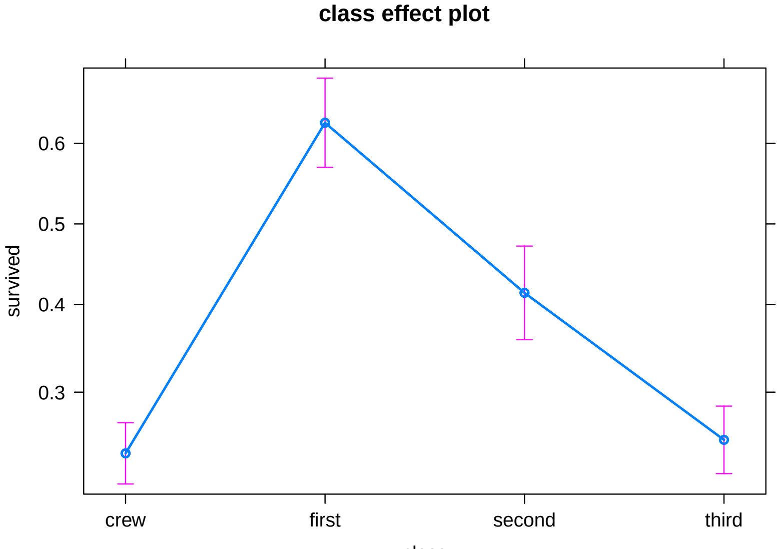
Presenting model results

```
kable(xtable::xtable(tit.glm), digits = 2)
```

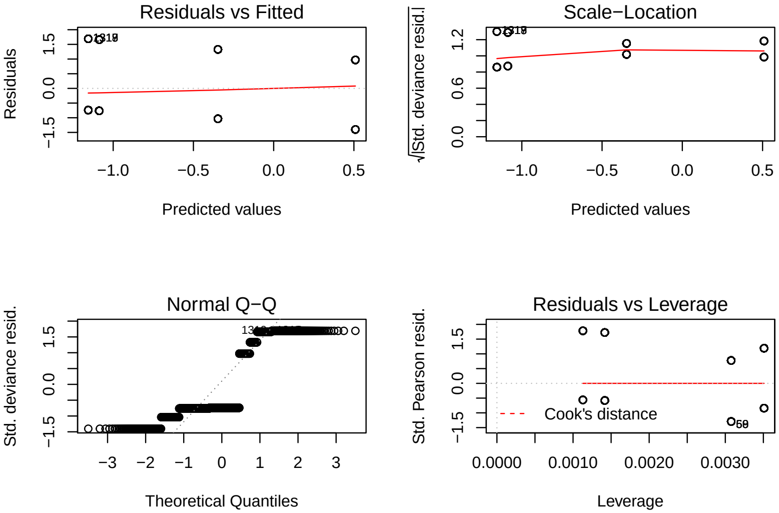
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.16	0.08	-14.67	0.00
classfirst	1.66	0.14	11.97	0.00
classsecond	0.81	0.14	5.62	0.00
classthird	0.07	0.12	0.58	0.56

Visualising model: `effects` package

```r
plot(allEffects(tit.glm))
```



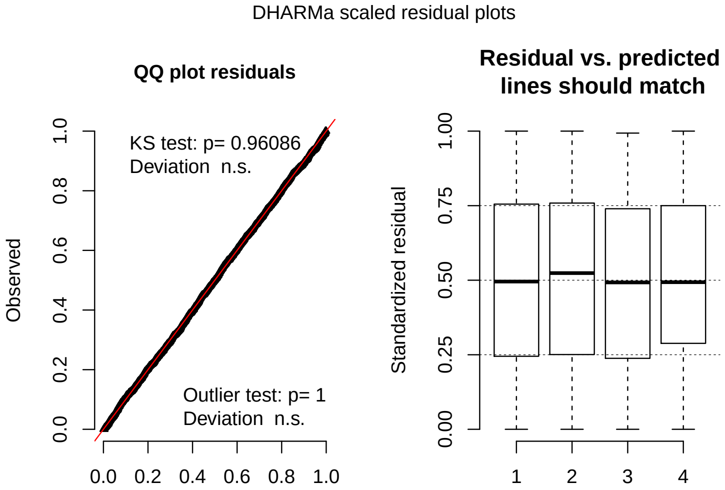
Logistic regression: model checking



```
null device
```

Residual diagnostics with DHARMa

```r
library(DHARMa)
simulateResiduals(tit.glm, plot = TRUE)
```



Pseudo R-squared for GLMs

```r
library(performance)
r2(tit.glm)
```

```
$R2_Tjur
 Tjur's R2
0.08650663
```

But many caveats apply! (e.g. see here and here)

Recapitulating

1. **Visualise data**

Recapitulating

1. **Visualise data**
2. **Fit model**: `glm`. Don't forget to specify `family`!

Recapitulating

1. **Visualise data**
2. **Fit model**: `glm`. Don't forget to specify `family`!
3. **Examine model**: `summary`

Recapitulating

1. **Visualise data**
2. **Fit model**: `glm`. Don't forget to specify `family`!
3. **Examine model**: `summary`
4. **Back-transform parameters** from *logit* into probability scale (e.g. `allEffects`)

Recapitulating

1. **Visualise data**
2. **Fit model**: `glm`. Don't forget to specify `family`!
3. **Examine model**: `summary`
4. **Back-transform parameters** from *logit* into probability scale (e.g. `allEffects`)
5. **Plot model**: `plot(allEffects(model))`, `visreg`, `plot_model`...

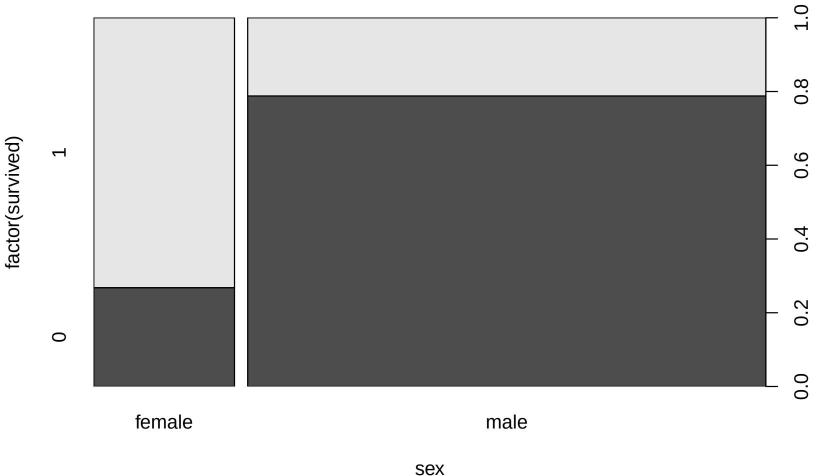
Recapitulating

1. **Visualise data**
2. **Fit model**: `glm`. Don't forget to specify `family`!
3. **Examine model**: `summary`
4. **Back-transform parameters** from *logit* into probability scale (e.g. `allEffects`)
5. **Plot model**: `plot(allEffects(model))`, `visreg`, `plot_model`...
6. **Examine residuals**: `DHARMa::simulateResiduals`.

Q: Did men have higher survival than women?

Plot first

```
plot(factor(survived) ~ sex, data = titanic)
```

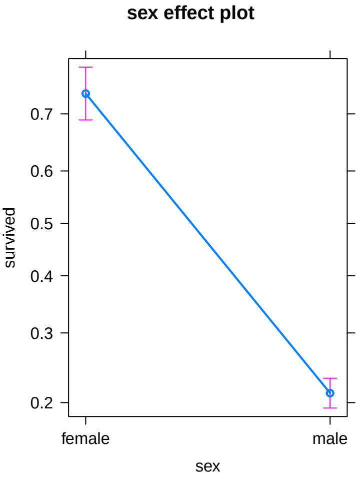


Fit model

```
Call:
glm(formula = survived ~ sex, family = binomial, data = titanic)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6226  -0.6903  -0.6903   0.7901   1.7613  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   1.0044     0.1041   9.645   <2e-16 ***
sexmale      -2.3172     0.1196 -19.376   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2335.0  on 2199  degrees of freedom
AIC: 2339
```

Effects

```
 model: survived ~ sex

 sex effect
sex
   female      male
0.7319149 0.2120162
```



Q: Did women have higher survival because they travelled more in first class?

Let's look at the data

```r
table(titanic$class, titanic$survived, titanic$sex)
```

```
, ,  = female


           0   1
  crew     3  20
  first    4 141
  second  13  93
  third  106  90

, ,  = male


           0   1
  crew   670 192
  first  118  62
  second 154  25
  third  422  88
```

Mmmm...

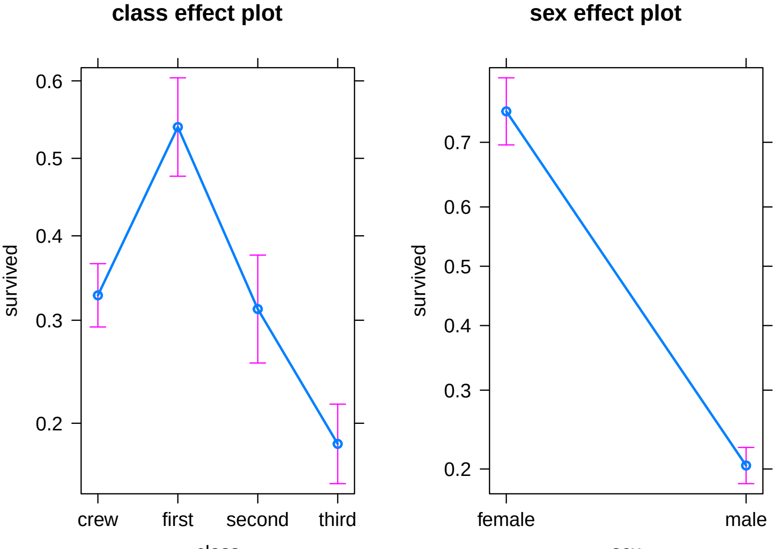
Fit additive model with both factors

```r
tit.sex.class <- glm(survived ~ class + sex, family = binomial,
```

```
glm(formula = survived ~ class + sex, family = binomial, data =
            coef.est coef.se
(Intercept)  1.19     0.16
classfirst   0.88     0.16
classsecond -0.07     0.17
classthird  -0.78     0.14
sexmale     -2.42     0.14
---
  n = 2201, k = 5
  residual deviance = 2228.9, null deviance = 2769.5 (difference
```

Plot additive model

```
plot(allEffects(tit.sex.class))
```



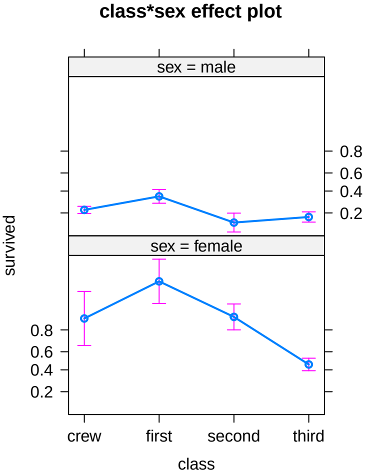
Fit model with both factors (interactions)

```
tit.sex.class <- glm(survived ~ class * sex, family = binomial,
```

```
glm(formula = survived ~ class * sex, family = binomial, data =
                     coef.est coef.se
(Intercept)           1.90     0.62
classfirst            1.67     0.80
classsecond           0.07     0.69
classthird           -2.06     0.64
sexmale              -3.15     0.62
classfirst:sexmale   -1.06     0.82
classsecond:sexmale -0.64     0.72
classthird:sexmale    1.74     0.65
---
  n = 2201, k = 8
  residual deviance = 2163.7, null deviance = 2769.5 (difference
```

Effects

```
model: survived ~ class * sex

 class*sex effect
       sex
class        female      male
  crew    0.8695652 0.2227378
  first   0.9724138 0.3444444
  second  0.8773585 0.1396648
  third   0.4591837 0.1725490
```



So, women had higher probability of survival than men, even within the same class.

Logistic regression for proportion data

Read Titanic data in different format

Read `Titanic_prop.csv` data.

```
  X Class    Sex   Age  No Yes
1 1   1st Female Adult   4 140
2 2   1st Female Child   0   1
3 3   1st   Male Adult 118  57
4 4   1st   Male Child   0   5
5 5   2nd Female Adult  13  80
6 6   2nd Female Child   0  13
```

These are the same data, but summarized (see `Freq` variable).

Use cbind(n.success, n.failures) as response

```r
prop.glm <- glm(cbind(Yes, No) ~ Class, data = tit.prop, family
```

```

Call:
glm(formula = cbind(Yes, No) ~ Class, family = binomial, data =

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.6404  -0.2915   1.5698   5.0366  10.1516

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5092     0.1146   4.445 8.79e-06 ***
Class2nd     -0.8565     0.1661  -5.157 2.51e-07 ***
Class3rd     -1.5965     0.1436 -11.114  < 2e-16 ***
ClassCrew    -1.6643     0.1390 -11.972  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Effects

```
 model: cbind(Yes, No) ~ Class

 Class effect
Class
      1st       2nd       3rd      Crew
0.6246154 0.4140351 0.2521246 0.2395480
```

Compare with former model based on raw data:

```
 model: survived ~ class

 class effect
class
     crew     first    second     third
0.2395480 0.6246154 0.4140351 0.2521246
```

Same results!

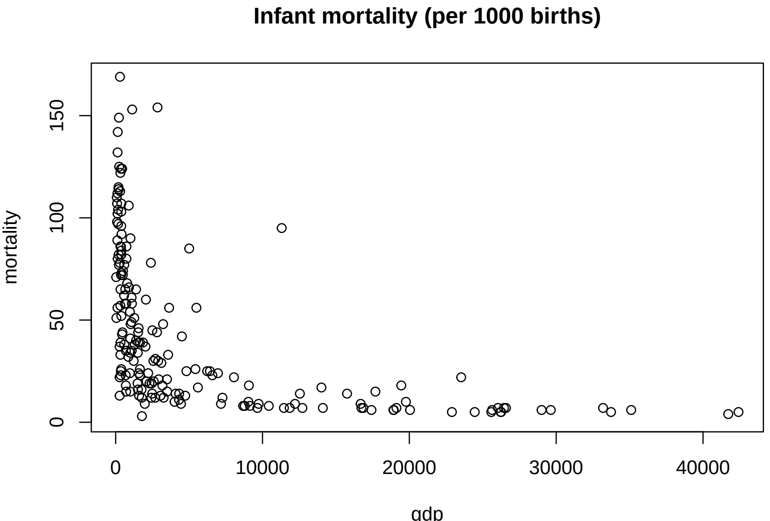
Logistic regression with continuous predictors

Example dataset: GDP and infant mortality
Read `UN_GDP_infantmortality.csv`.

```
          country          mortality           gdp
Afghanistan   :  1   Min.   :  2.00   Min.   :   36
Albania       :  1   1st Qu.: 12.00   1st Qu.:  442
Algeria       :  1   Median : 30.00   Median : 1779
American.Samoa:  1   Mean   : 43.48   Mean   : 6262
Andorra       :  1   3rd Qu.: 66.00   3rd Qu.: 7272
Angola        :  1   Max.   :169.00   Max.   :42416
(Other)       :201   NA's   :6        NA's   :10
```

EDA

```
plot(mortality ~ gdp, data = gdp, main = "Infant mortality (per
```



Fit model

```r
gdp.glm <- glm(cbind(mortality, 1000 - mortality) ~ gdp,
               data = gdp, family = binomial)
```

```
Call:
glm(formula = cbind(mortality, 1000 - mortality) ~ gdp, family =
    data = gdp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.2230  -3.5163  -0.5697   2.4284  13.5849

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.657e+00  1.311e-02 -202.76   <2e-16 ***
gdp         -1.279e-04  3.458e-06  -36.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

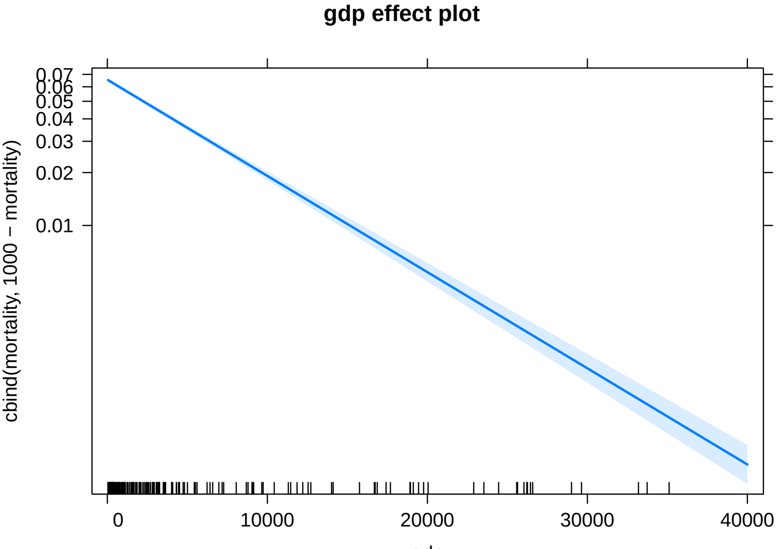
Effects

```r
allEffects(gdp.glm)
```

```
 model: cbind(mortality, 1000 - mortality) ~ gdp

 gdp effect
gdp
          40        10000        20000        30000        40000 
0.0652177296 0.0191438829 0.0054028095 0.0015096074 0.0004206154 
```

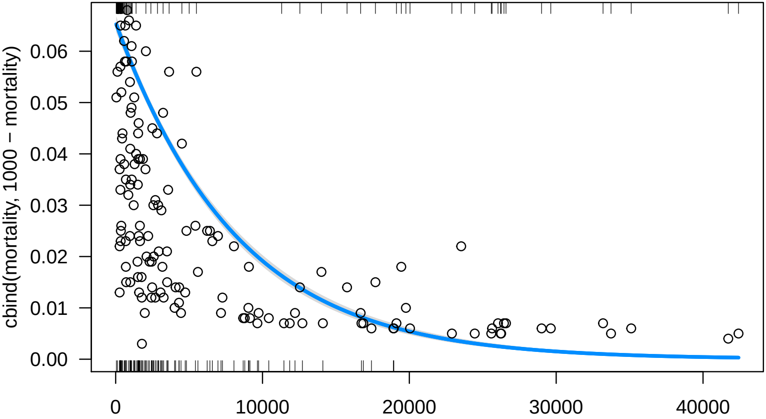
Effects plot

```
plot(allEffects(gdp.glm))
```



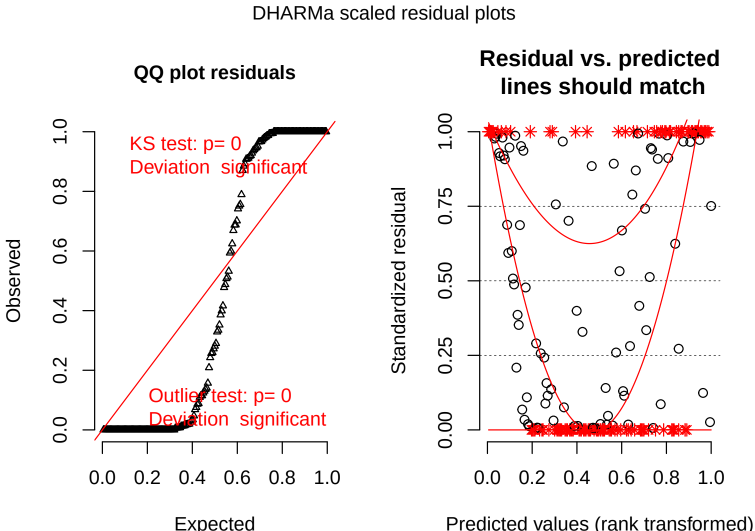
Plot model using visreg:

```
visreg(gdp.glm, scale = "response")
points(mortality/1000 ~ gdp, data = gdp)
```



Residuals diagnostics with DHARMa

```
simulateResiduals(gdp.glm, plot = TRUE)
```



Overdispersion

Testing for overdispersion (DHARMa)

```r
simres <- simulateResiduals(gdp.glm, refit = TRUE)
testDispersion(simres, plot = FALSE)
```

```

	DHARMa nonparametric dispersion test via mean deviance resid
	fitted vs. simulated-refitted

data:  simres
dispersion = 21, p-value < 2.2e-16
alternative hypothesis: two.sided
```

Overdispersion in logistic regression with proportion data

```
gdp.overdisp <- glm(cbind(mortality, 1000 - mortality) ~ gdp,
                    data = gdp, family = quasibinomial)
```

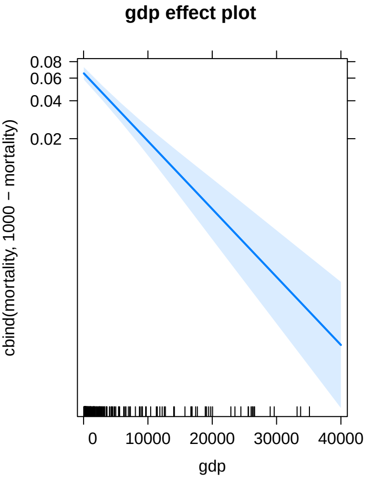
```
Call:
glm(formula = cbind(mortality, 1000 - mortality) ~ gdp, family =
    data = gdp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.2230  -3.5163  -0.5697   2.4284  13.5849

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.657e+00  5.977e-02 -44.465  < 2e-16 ***
gdp         -1.279e-04  1.577e-05  -8.111 5.96e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 20.79
```

Mean estimates do not change after accounting for overdispersion

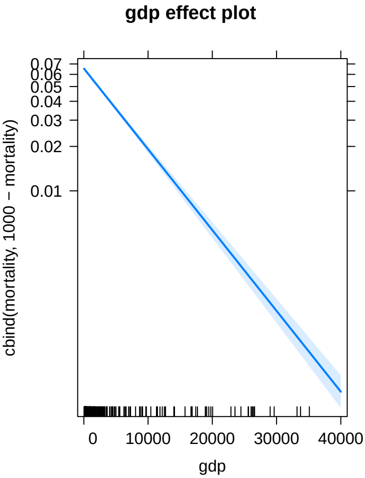
```
 model: cbind(mortality, 1000 - mortality) ~ gdp

 gdp effect
gdp
          40        10000        20000        30000        40000
0.0652177296 0.0191438829 0.0054028095 0.0015096074 0.0004206154
```

```
 model: cbind(mortality, 1000 - mortality) ~ gdp

 gdp effect
gdp
          40        10000        20000        30000        40000
0.0652177296 0.0191438829 0.0054028095 0.0015096074 0.0004206154
```

But standard errors (uncertainty) do!

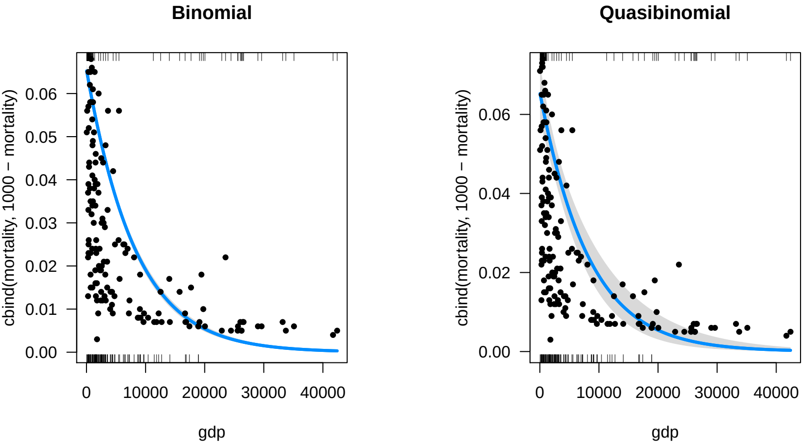


gdp effect plot



gdp effect plot

Plot model and data

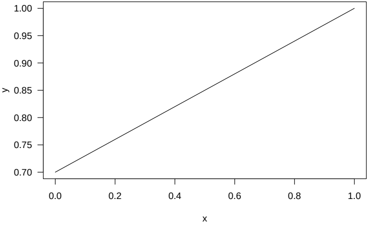


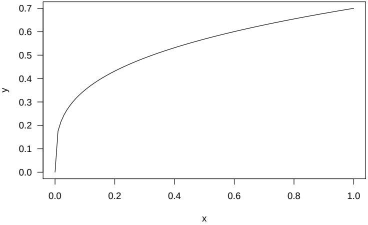
Overdispersion

Whenever you fit logistic regression to **proportion** data, check family `quasibinomial`.

Think about the shape of relationships

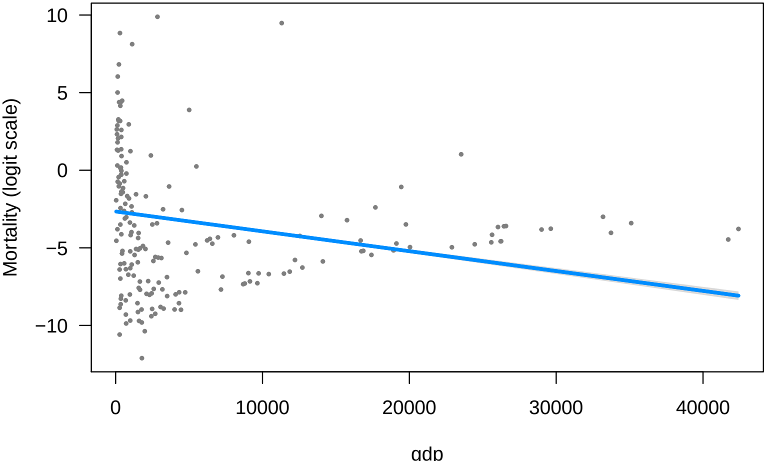
$y \sim x + z$
Really? Not everything has to be linear! Actually, it often is not.
Think about shape of relationship. See chapter 3 in Bolker's book.



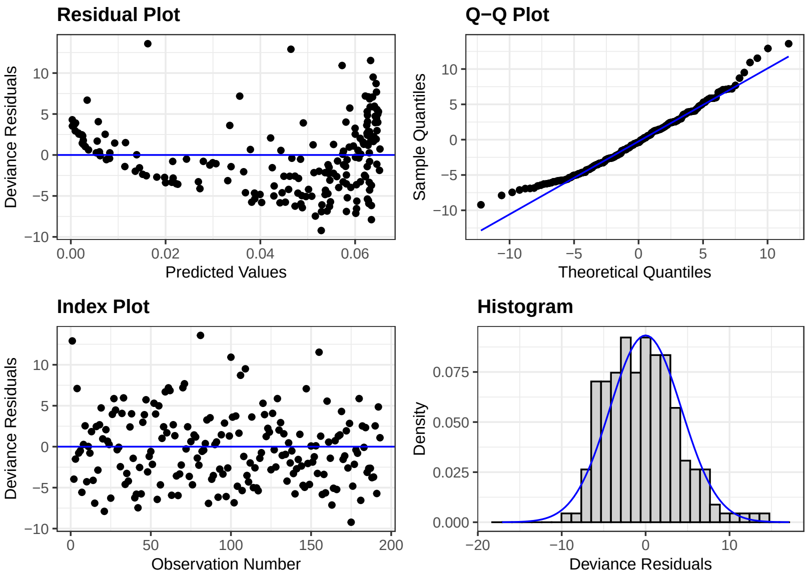


Think about the shape of relationships

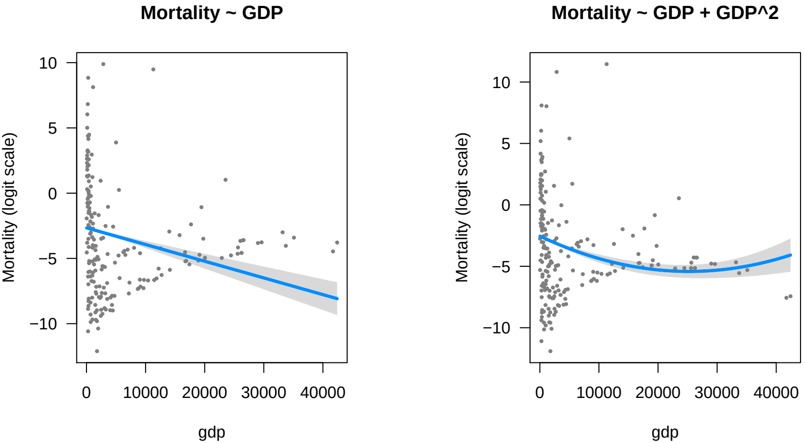
```
visreg(gdp.glm, ylab = "Mortality (logit scale)")
```



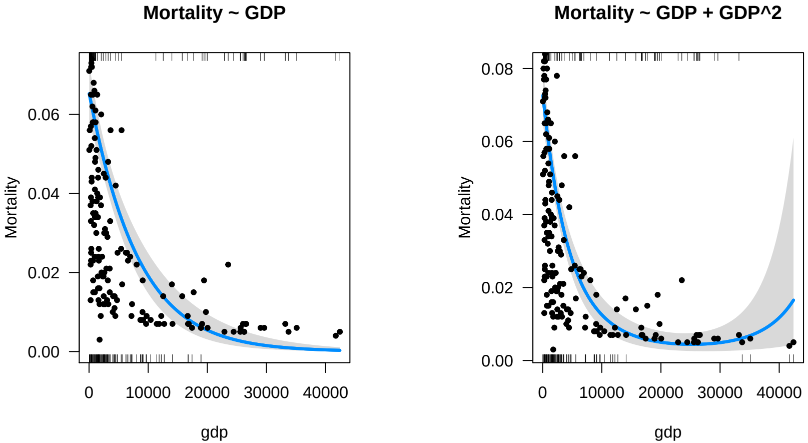
Think about the shape of relationships



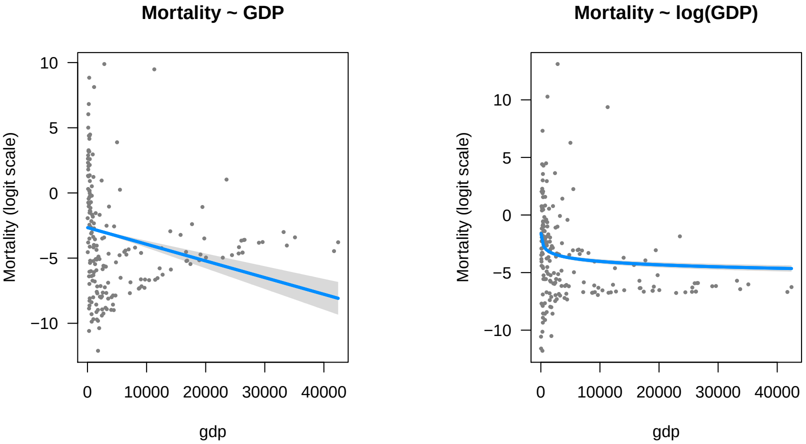
Think about the shape of relationships



Think about the shape of relationships



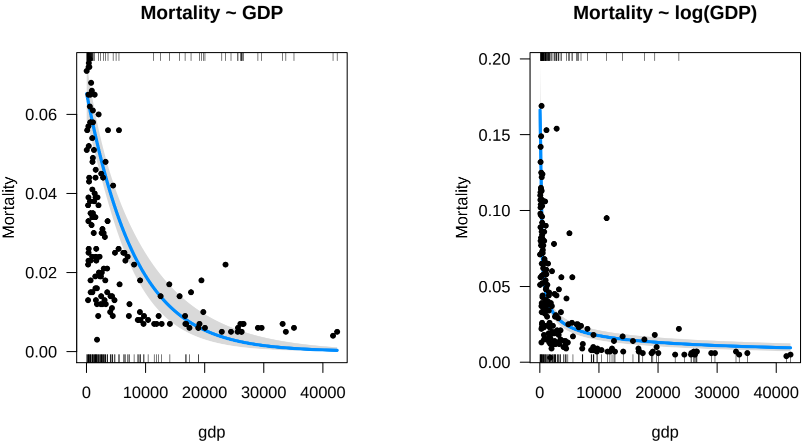
Think about the shape of relationships



Mortality ~ GDP

Mortality ~ log(GDP)

Think about the shape of relationships



More examples

- trees.csv: probability of tree death in relation to size

More examples

- trees.csv: probability of tree death in relation to size
- seedset.csv: Comparing seed set among plants (Data from Harder et al. 2011)

Seed set among plants

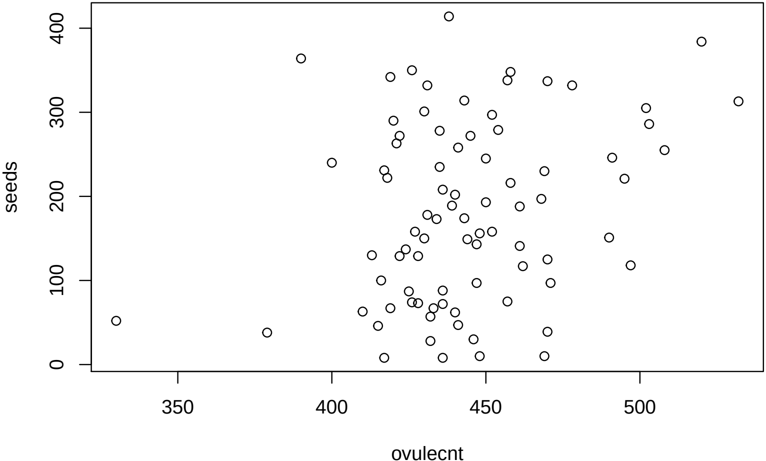
```
seed <- readr::read_csv("data-raw/seedset.csv")
head(seed)
```

```
# A tibble: 6 x 6
  species    plant pcmass fertilized seeds ovulecnt
  <chr>      <dbl>  <dbl>      <dbl> <dbl>    <dbl>
1 ferruginea     2  0             70    52      330
2 ferruginea     2  0.2          321   188      461
3 ferruginea     2  0.485        351   278      435
4 ferruginea     2  0.737        386   301      430
5 ferruginea     2  1            367   342      419
6 ferruginea     3  0            185    39      470
```

```
seed$plant <- as.factor(seed$plant)
```

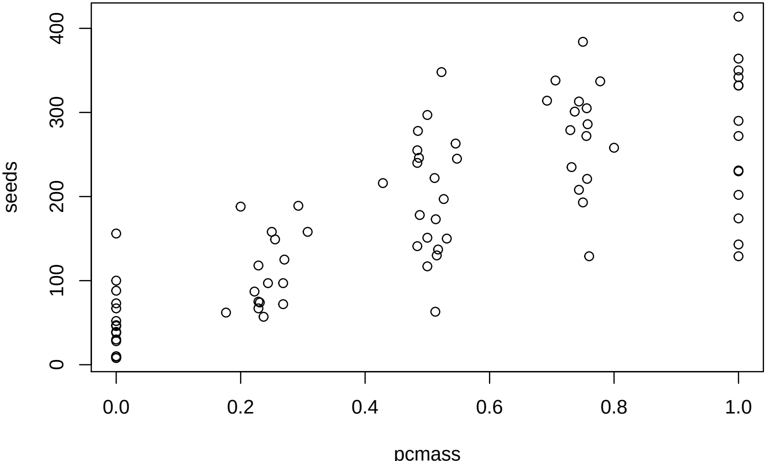
Number of seeds vs Number of ovules

```r
plot(seeds ~ ovulecnt, data = seed)
```

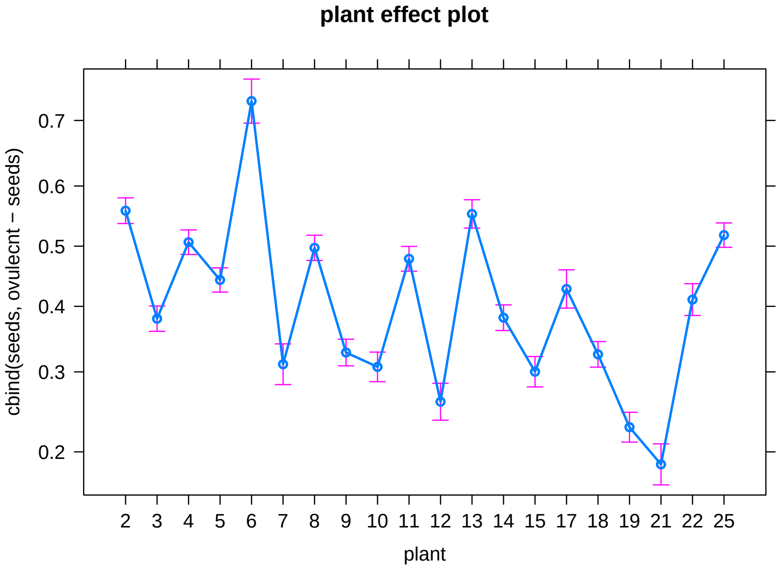


Number of seeds vs Proportion outcross pollen

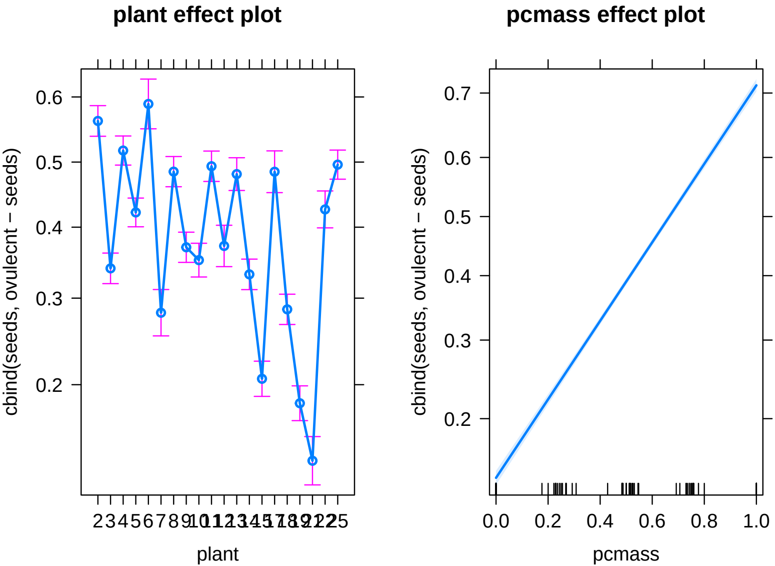
```
plot(seeds ~ pcmass, data = seed)
```



Seed set across plants



Seed set ∼ outcross pollen



GLM for count data: Poisson regression

Types of response variable

- Gaussian: `lm`

Types of response variable

- Gaussian: `lm`
- Bernouilli / Binomial: `glm` (family `binomial / quasibinomial`)

Types of response variable

- Gaussian: `lm`
- Bernouilli / Binomial: `glm` (family `binomial / quasibinomial`)
- Counts: `glm` (family `poisson / quasipoisson`)

Poisson regression

- Response variable: Counts (0, 1, 2, 3…) - discrete

Then

$$log(N) = a + bx$$
$$N = e^{a+bx}$$

Poisson regression

- Response variable: Counts (0, 1, 2, 3…) - discrete
- Link function: `log`

Then

$$log(N) = a + bx$$
$$N = e^{a+bx}$$

Example dataset: Seedling counts in quadrats

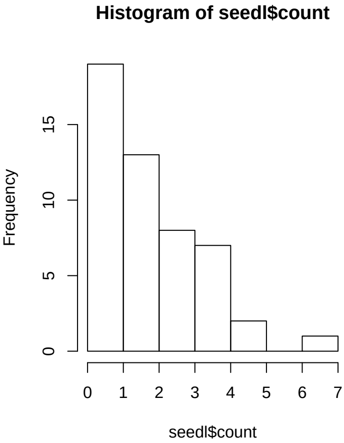
```r
seedl <- read.csv("data-raw/seedlings.csv")
```

```
       X             count          row         col      
 Min.   : 1.00   Min.   :0.00   Min.   :1   Min.   : 1.0  
 1st Qu.:13.25   1st Qu.:1.00   1st Qu.:2   1st Qu.: 3.0  
 Median :25.50   Median :2.00   Median :3   Median : 5.5  
 Mean   :25.50   Mean   :2.14   Mean   :3   Mean   : 5.5  
 3rd Qu.:37.75   3rd Qu.:3.00   3rd Qu.:4   3rd Qu.: 8.0  
 Max.   :50.00   Max.   :7.00   Max.   :5   Max.   :10.0  
     light             area     
 Min.   : 2.571   Min.   :0.25  
 1st Qu.:26.879   1st Qu.:0.25  
 Median :47.493   Median :0.50  
 Mean   :47.959   Mean   :0.62  
 3rd Qu.:67.522   3rd Qu.:1.00  
 Max.   :99.135   Max.   :1.00  
```

EDA

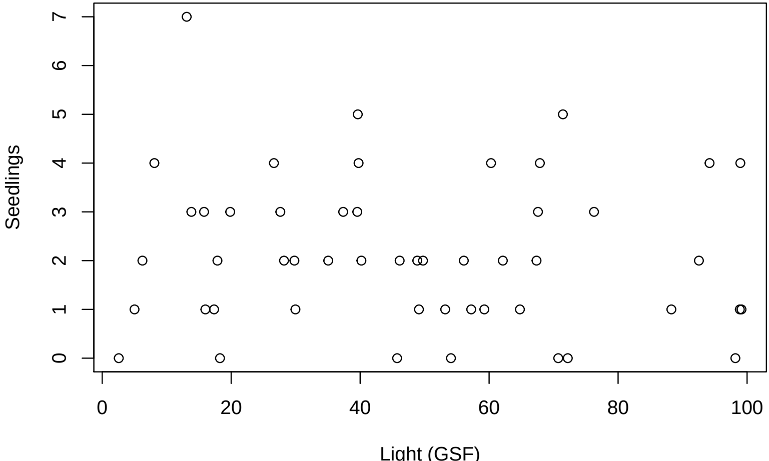
```r
table(seedl$count)
```

```
 0  1  2  3  4  5  7
 7 12 13  8  7  2  1
```



Q: Relationship between Nseedlings and light?

```
plot(seedl$light, seedl$count, xlab = "Light (GSF)", ylab = "See
```



Let's fit model (Poisson regression)

```r
seedl.glm <- glm(count ~ light, data = seedl, family = poisson)
summary(seedl.glm)
```

```
Call:
glm(formula = count ~ light, family = poisson, data = seedl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1906  -0.8466  -0.1110   0.5220   2.4577

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.881805   0.188892   4.668 3.04e-06 ***
light       -0.002576   0.003528  -0.730    0.465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 63.029  on 49  degrees of freedom
Residual deviance: 62.492  on 48  degrees of freedom
AIC: 182.03

Number of Fisher Scoring iterations: 5
```

Interpreting Poisson regression output

Parameter estimates (log scale):

```r
coef(seedl.glm)
```

```
 (Intercept)        light 
 0.881805022 -0.002575656 
```

We need to back-transform: apply the inverse of the logarithm

```r
exp(coef(seedl.glm))
```

```
(Intercept)       light 
  2.4152554   0.9974277 
```

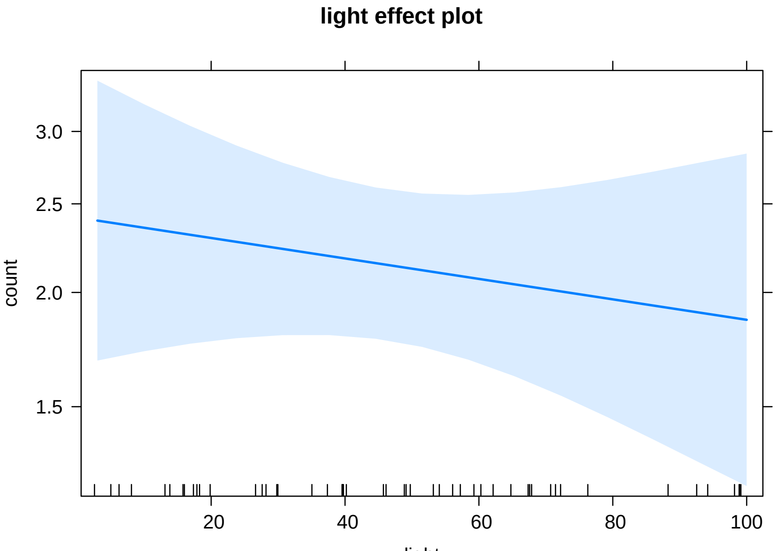
Using effects package

```r
summary(allEffects(seedl.glm))
```

```
 model: count ~ light

 light effect
light
       3       30       50       70      100
2.396665 2.235657 2.123408 2.016794 1.866826

 Lower 95 Percent Confidence Limits
light
       3       30       50       70      100
1.684579 1.795202 1.753373 1.567785 1.228247

 Upper 95 Percent Confidence Limits
light
       3       30       50       70      100
3.409754 2.784179 2.571535 2.594398 2.837408
```

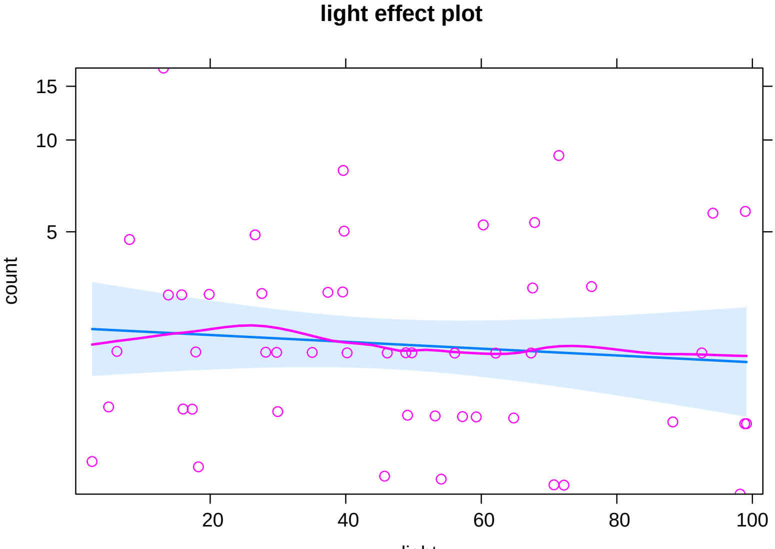
So what's the relationship between Nseedlings and light?

```
plot(allEffects(seedl.glm))
```

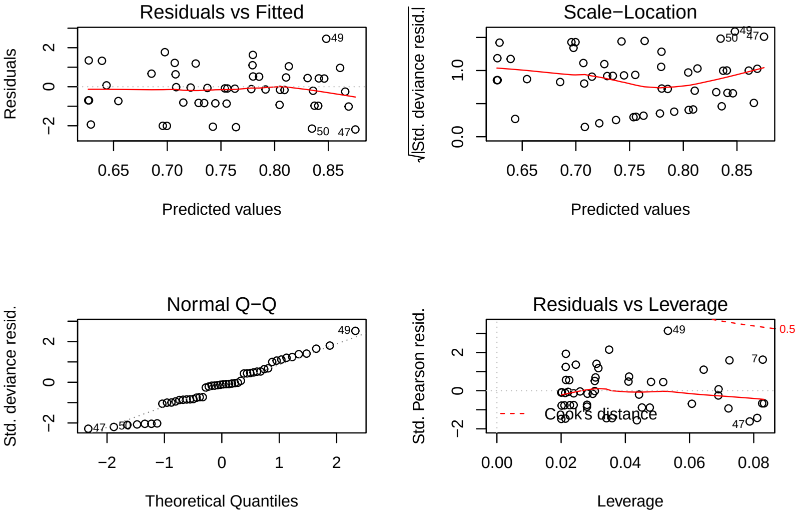


There's lot of unexplained variation

```
plot(allEffects(seedl.glm, residuals = TRUE))
```



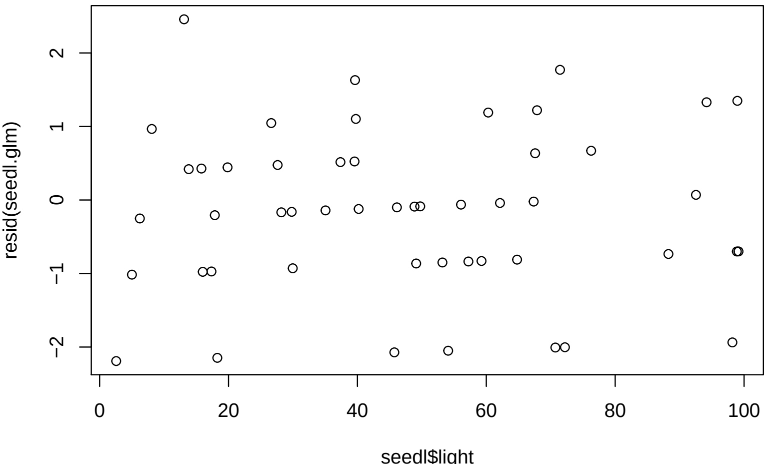
Poisson regression: model checking



```
null device
```

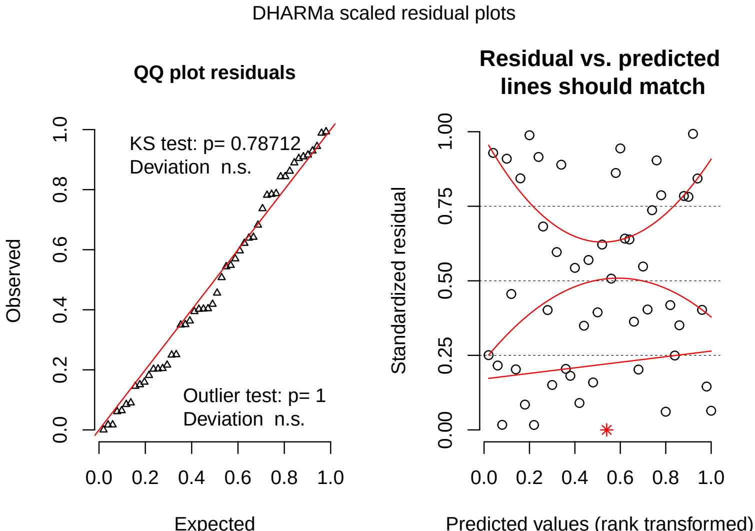
Is there pattern of residuals along predictor?

```r
plot(seedl$light, resid(seedl.glm))
```



Residuals diagnostics with DHARMa

```
DHARMa::simulateResiduals(seedl.glm, plot = TRUE)
```



Poisson regression: Overdispersion

Always check overdispersion with count data

```r
simres <- simulateResiduals(seedl.glm, refit = TRUE)
testDispersion(simres, plot = FALSE)
```

```

	DHARMa nonparametric dispersion test via mean deviance resid
	fitted vs. simulated-refitted

data:  simres
dispersion = 1.1655, p-value = 0.432
alternative hypothesis: two.sided
```

Accounting for overdispersion in count data

Use family quasipoisson

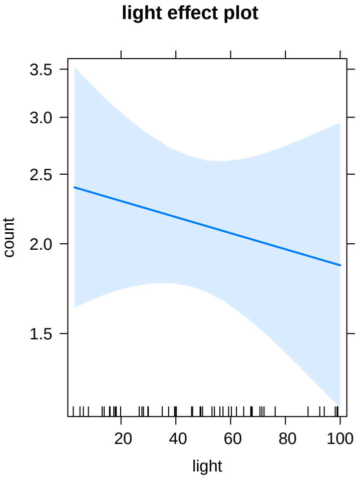
```

Call:
glm(formula = count ~ light, family = quasipoisson, data = seedl

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1906  -0.8466  -0.1110   0.5220   2.4577

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.881805   0.201230   4.382 6.37e-05 ***
light       -0.002576   0.003758  -0.685    0.496
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.1349

    Null deviance: 63.029  on 49  degrees of freedom
Residual deviance: 62.492  on 48  degrees of freedom
AIC: NA
```

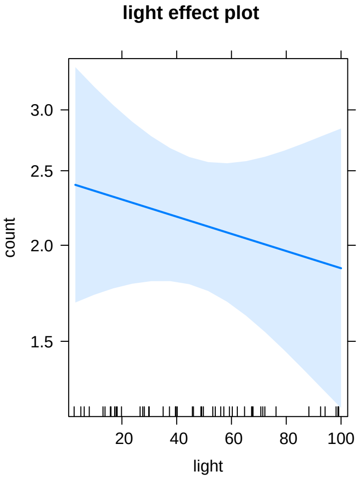
Mean estimates do not change after accounting for overdispersion

```
 model: count ~ light

 light effect
light
       3       30       50       70      100 
2.396665 2.235657 2.123408 2.016794 1.866826 
```

```
 model: count ~ light

 light effect
light
       3       30       50       70      100 
2.396665 2.235657 2.123408 2.016794 1.866826 
```

But standard errors may change

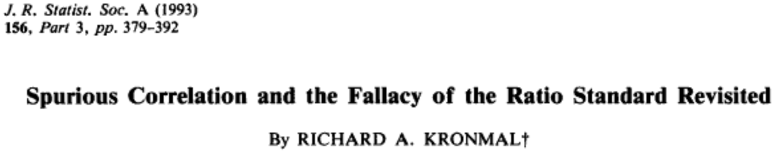




What if survey plots have different area?

Avoid regression of ratios

seedlings/area $\sim$ light



J. R. Statist. Soc. A (1993)
156, *Part* 3, *pp*. 379–392

Spurious Correlation and the Fallacy of the Ratio Standard Revisited

By RICHARD A. KRONMAL†

Use offset to standardise response variables in GLMs

```r
seedl.offset <- glm(count ~ light, offset = seedl$area, data = s
summary(seedl.offset)
```

```
Call:
glm(formula = count ~ light, family = poisson, data = seedl,
    offset = seedl$area)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6926  -0.8532   0.1491   0.5211   3.1051

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.299469   0.185468   1.615    0.106
light       -0.004498   0.003441  -1.307    0.191

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70.263  on 49  degrees of freedom
```

Note estimates now referred to area units

```
exp(coef(seedl.offset))
```

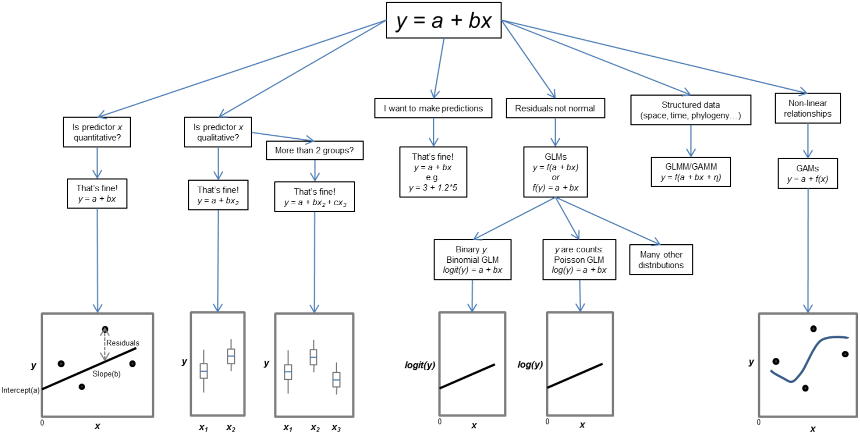
```
(Intercept)       light
  1.3491422   0.9955123
```

Other examples

- Infant mortality $\sim$ GDP

Other examples

- Infant mortality $\sim$ GDP
- Number of cones consumed by squirrels (data)



$y = a + bx$

Is predictor x quantitative?

That's fine! $y = a + bx$

Is predictor x qualitative?

That's fine! $y = a + bx_2$

More than 2 groups?

That's fine! $y = a + bx_2 + cx_3$

I want to make predictions

That's fine! $y = a + bx$ e.g. $y = 3 + 1.2*5$

Residuals not normal

GLMs $y = f(a + bx)$ *or* $f(y) = a + bx$

Binary y: Binomial GLM $logit(y) = a + bx$

y are counts: Poisson GLM $log(y) = a + bx$

Many other distributions

Structured data (space, time, phylogeny…)

GLMM/GAMM $y = f(a + bx + \eta)$

Non-linear relationships

GAMs $y = a + f(x)$

END

cc BY NC SA

Source code and materials:
https://github.com/Pakillo/LM-GLM-GLMM-intro