ENV 710: Lecture 10

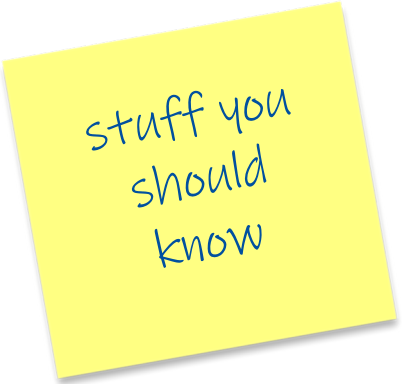
linear models

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

linear models

general linear models

learning goals



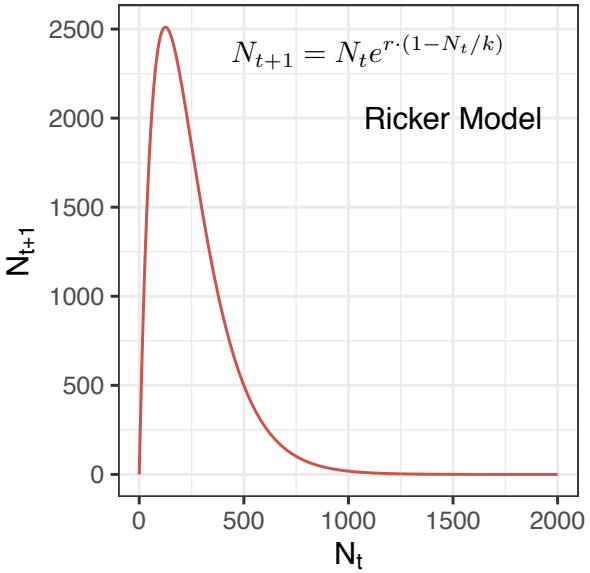
- what are models and linear models?
- models with continuous response and explanatory variables – simple linear regression
- parameter estimation
- analysis of variance for partitioning variance into different components and testing model fit to the data
- coefficient of determination, R^2
- simple linear modeling in R
- model validation techniques

what are models?

models

- models simplify and explain variability in the natural world
- models are representations of reality, should be as accurate and convenient as possible
- some models are mechanistic
- some models are empirical (based on the data themselves) and used when the underlying mechanism is unknown

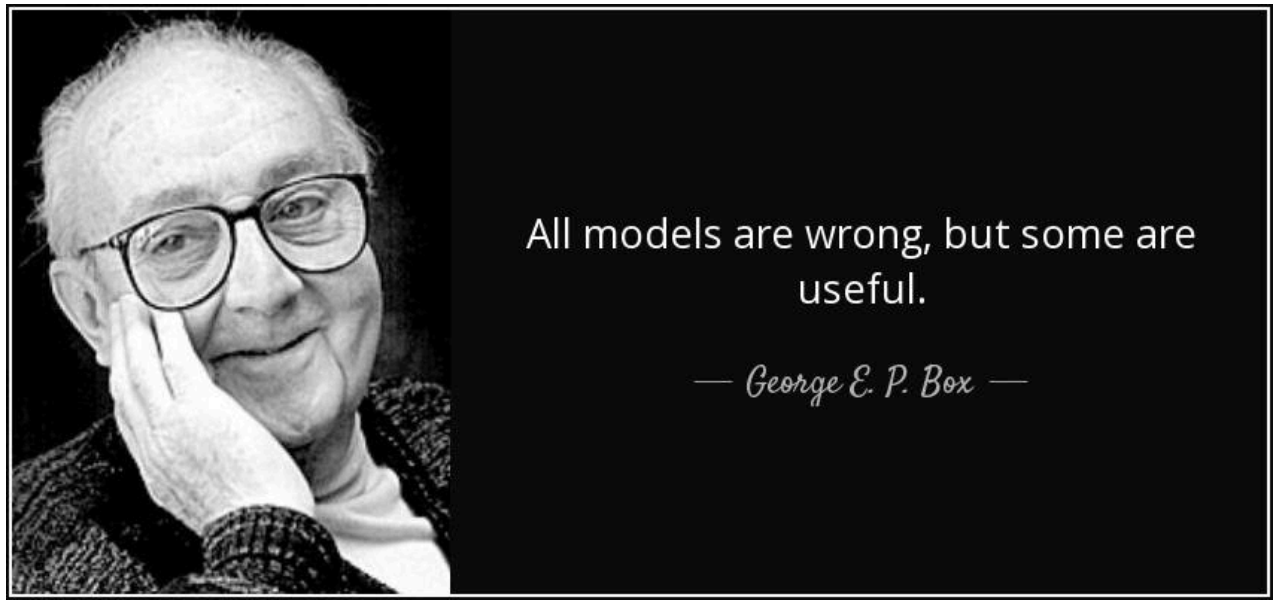
discrete population model



what are models?

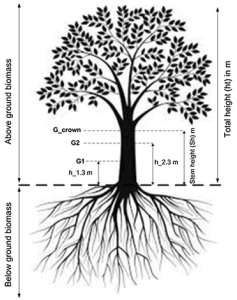
- many potential models
- most modeling is a tradeoff between maximizing a model's realism and usefulness
- goal is to find the minimum adequate model, or the model that corresponds to the design of your experiment
- simpler models are often preferred to complex models when there is no reduction in explanatory power

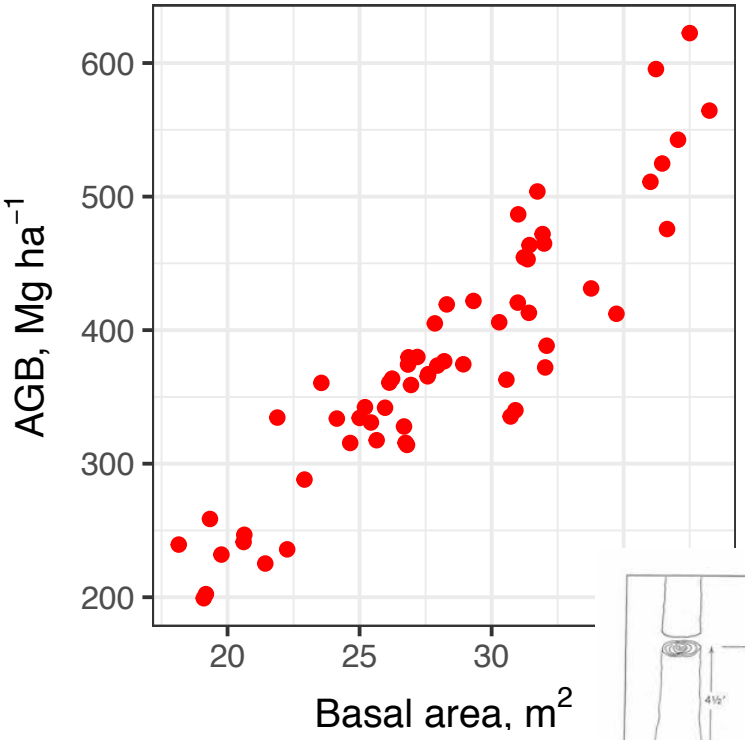
- modelling philosophy will depend on your goals:
  - for planned experiments, fit the model that represents your experiment
  - for observational studies, fit the maximal model and simplify to achieve a minimum adequate model
  - model simplification can be done using different procedures… more to come
- no fixed rules and no absolutes
- modeling should include your environmental knowledge, and not be a purely mechanical process

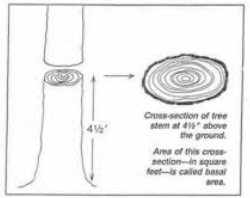

All models are wrong, but some are useful.
— George E. P. Box —

what are linear models?

- investigate whether areas with bigger trees, i.e. higher basal area, result in forest stands with higher biomass
- what is the simplest model to fit these data?
- how would you fit a line to the data by eye?





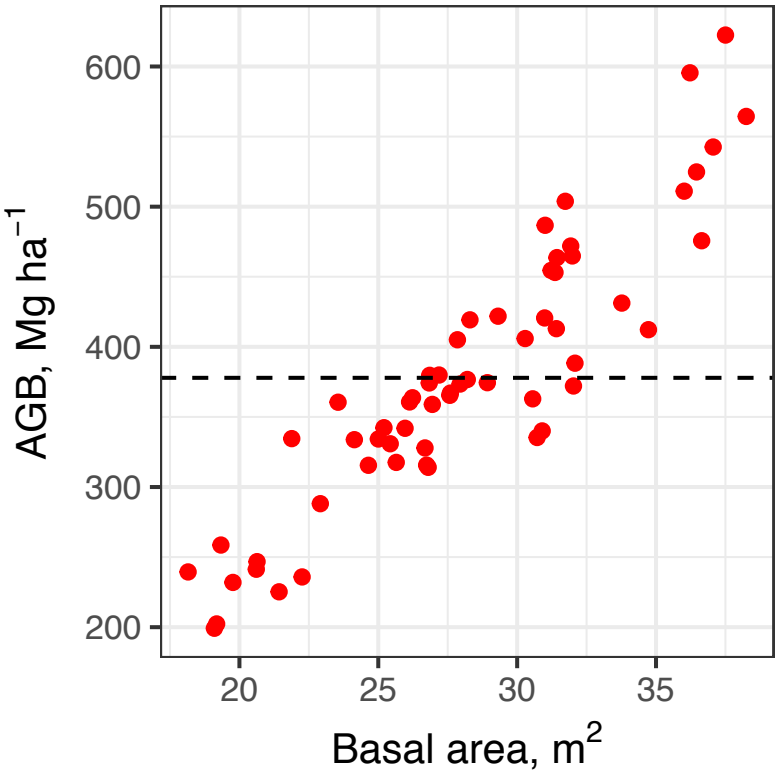


what are linear models?

- investigate whether areas with bigger trees, i.e. higher basal area, result in forest stands with higher biomass
- what is the simplest model to fit these data?

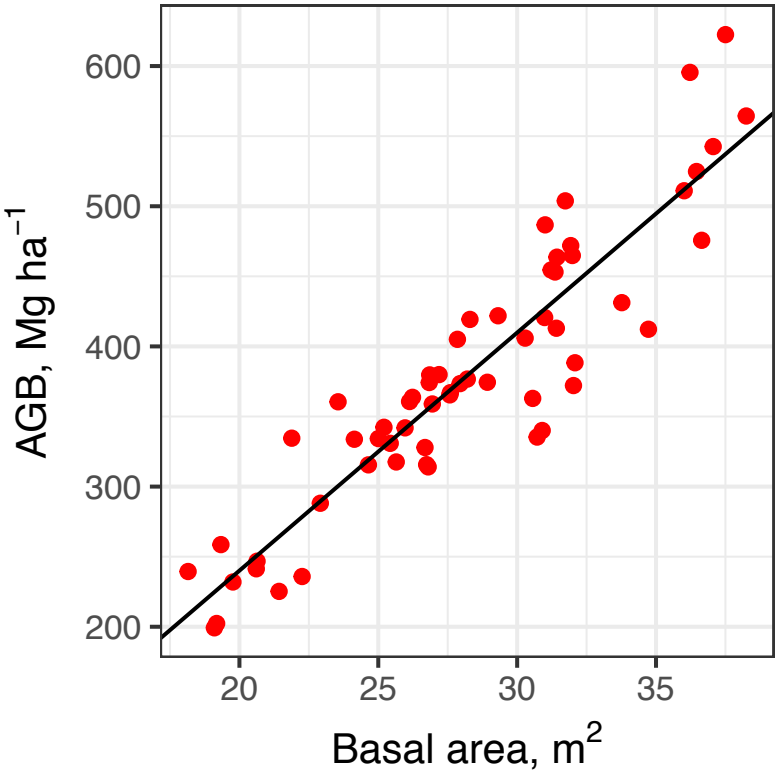
$$y = a$$

- the relationship between basal area and AGB is constant → null model



what are linear models?

- next simplest model is a linear relationship between basal area and AGB
- AGB is hypothesized to be a linear function of basal area
- rather than fit this by eye, we can define the relationship mathematically

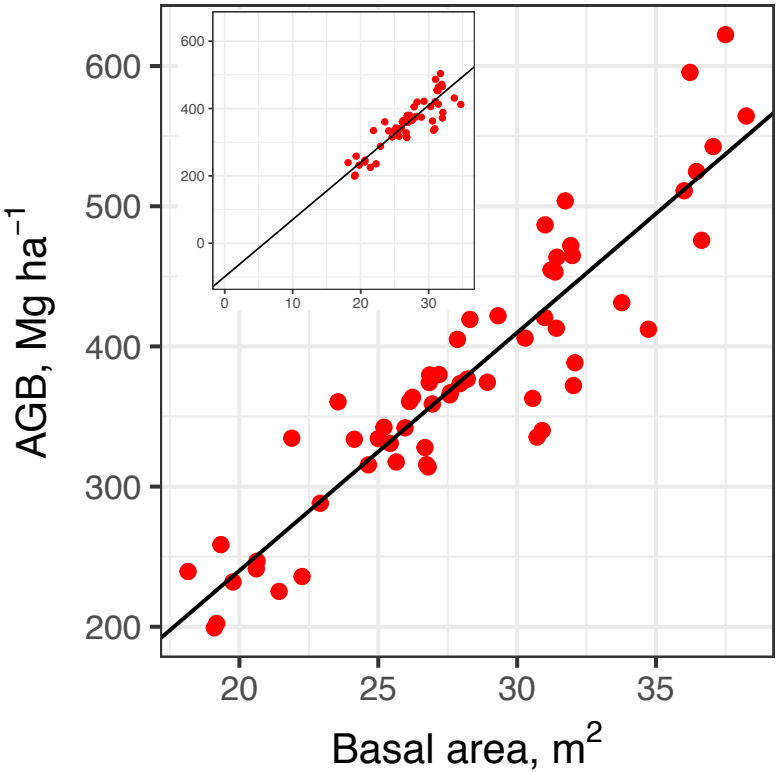


what are linear models?

- equation of a straight line

$$y = a + bx$$

- equation is fit by 2 parameters, *a* and *b*
- *a* = intercept = -99



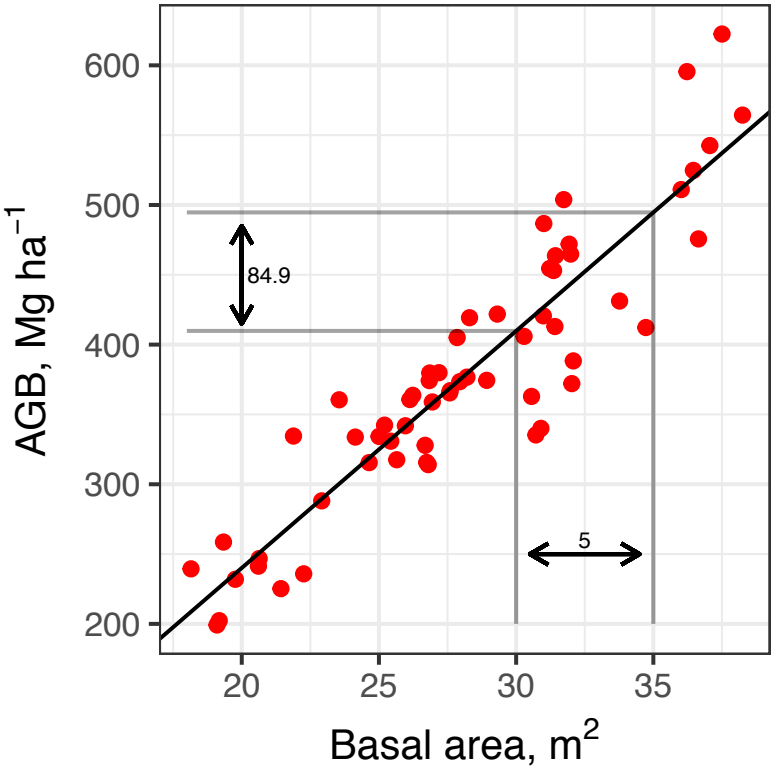
what are linear models?

- equation of a straight line

$$y = a + bx$$

- equation is fit by 2 parameters, *a* and *b*
- *b* = slope = 16.98

$$b = \frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{494.7 - 409.8}{35 - 30} = 16.98$$



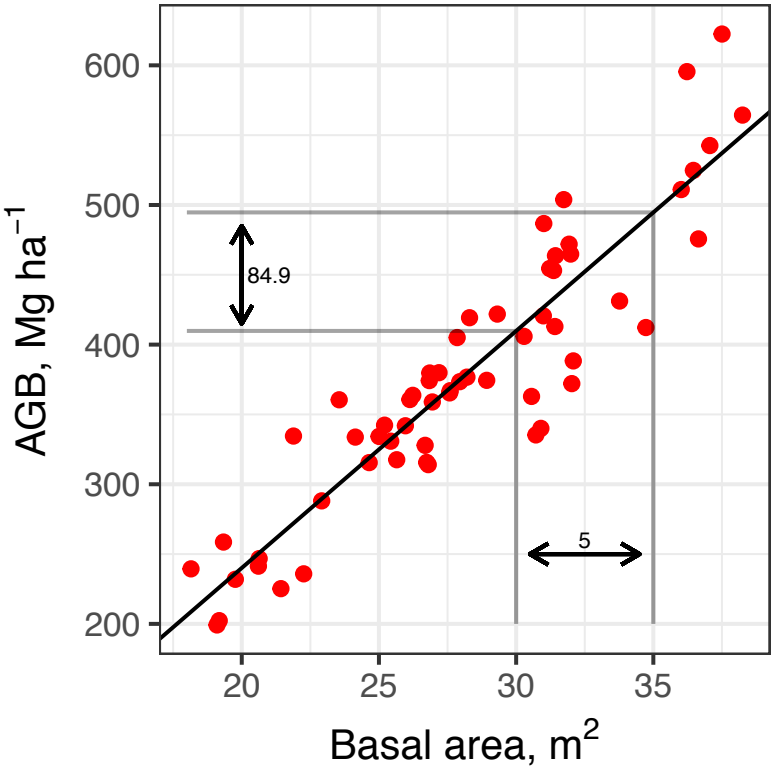
what are linear models?

- equation of the linear relationship between basal area and biomass is:

$$AGB = -99 + 16.98(BA)$$

- placed line to minimize the distance of points to the line - "line of best fit"

- line describes the relationship that explains the maximum amount of variability



linear models

- what makes $y = a + bx$ a linear model?
  - parameters combine additively
  - any model that has parameters (ie. a and b) separated by a + or – sign is linear

    $y = a + bx_1 + cx_2 + dx_3$ $\qquad$ $y = a + be^x$
  - common misconception that linear models cannot be fitted to curved data

    $y = a + b\sin(x)$
  - the model needs to be linear, not necessarily the relationship(s)

    $y = a + bx + cx^2$

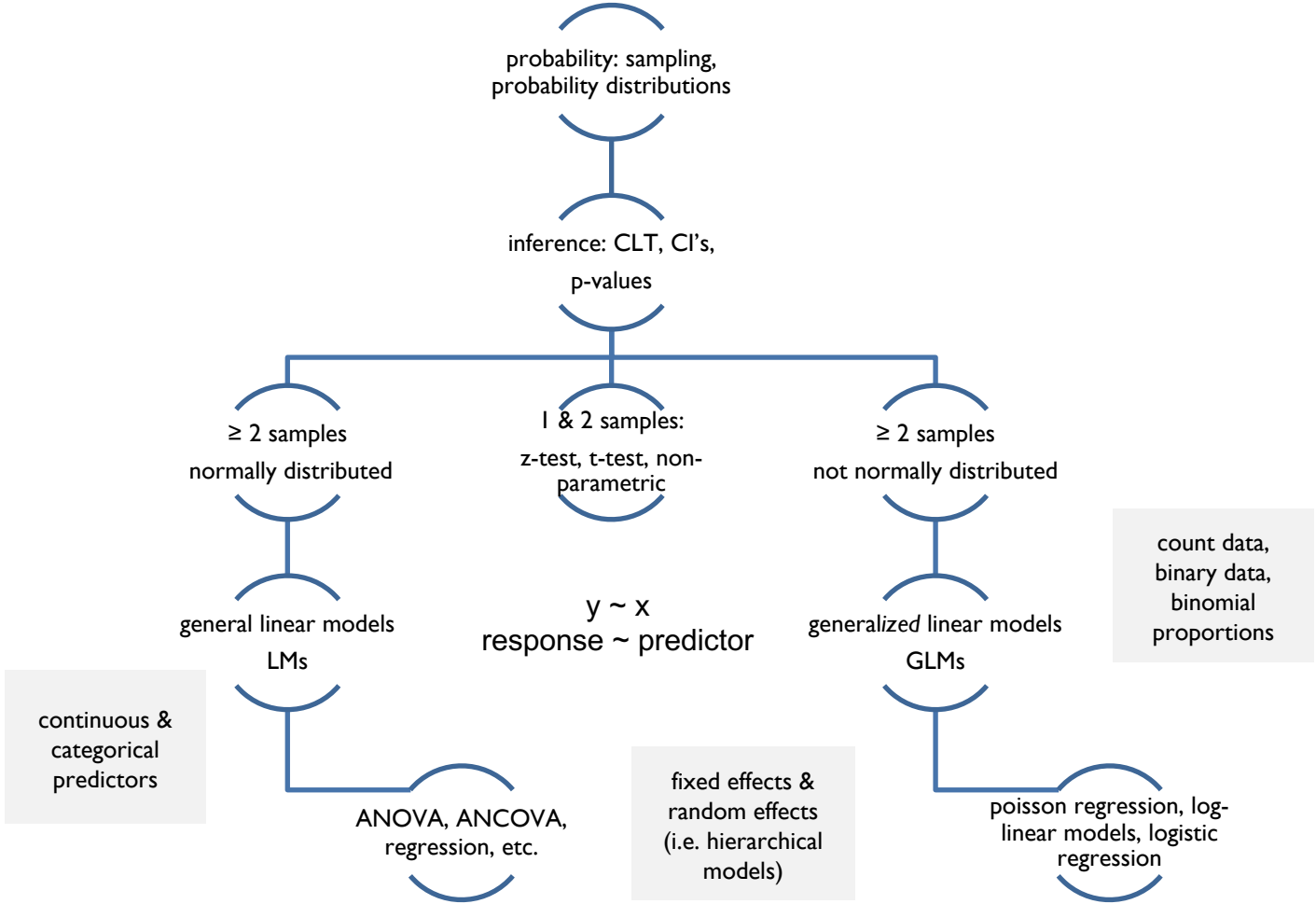
what are *general linear models*?

- traditionally, statistics is performed (and taught) using a recipe book (t-test, ANOVA, ANCOVA)

- if you understand the fundamentals of general linear modelling, the rest of this course extends the concepts to other probability distributions

- general linear models provide a coherent and theoretically satisfying framework on which to conduct analyses

- parametric (i.e. normally distributed) general linear models
  - one sample t-test
  - two sample t-test
  - paired t-test
  - ANOVA
  - ANCOVA
  - correlation
  - linear regression
  - multiple regression


probability: sampling, probability distributions
inference: CLT, CI's, p-values
≥ 2 samples normally distributed
1 & 2 samples: z-test, t-test, non-parametric
≥ 2 samples not normally distributed
general linear models LMs
y ~ x
response ~ predictor
generalized linear models GLMs
count data, binary data, binomial proportions
continuous & categorical predictors
ANOVA, ANCOVA, regression, etc.
fixed effects & random effects (i.e. hierarchical models)
poisson regression, log-linear models, logistic regression

model formulae

- general linear modeling is based on the concept of model formulae

response variable ~ *explanatory variable(s)* + *error*

- read as 'variation in response variable modeled as a function of the explanatory variable(s) plus variation not explained by explanatory variable(s)'
- attributes of the response and explanatory variables determine the type of linear model that is fitted

y~*x*	if *y* and *x* are continuous then simple linear regression
	e.g., seed mass (g) ~ seed length (mm)
y~*factor(x)*	if *x* is a categorical (nominal) variable then one-way ANOVA
	e.g., seed mass (g) ~ dispersal type (wind, animal, water)

linear model formulae

model formula	traditional name
$y \sim x_1$ (continuous)	regression
$y \sim x_1$ (categorical)	one-way ANOVA
$y \sim x_1$ (cat) + x_2 (cat)	two-way ANOVA
$y \sim x_1$ (cat) + x_2 (cont)	ANCOVA
$y \sim x_1$ (cont) + X_2 (cont)	multiple regression
$y \sim x_1$ (cat) • X_2 (cat)	factorial ANOVA

LET'S RECAP...

- models are a trade-off between realism and interpretability
- maximize the explained variance in the response with the simplest model
- we (often) start with the maximal model and simplify to the minimum adequate model
- general linear models provide a powerful and flexible framework to analyze a variety of data and experimental designs
- general linear models include 'traditional' statistical techniques
- general linear models are based around model formulae

linear models

one continuous IV

simple linear regression

- simple linear regression is used when both the response and explanatory variables are continuous
  - response variable is assumed to be random (stochastic) and therefore has a probability distribution
  - explanatory variables are assumed to have fixed values
- regression can quantify the strength between Y and X
- regression can be used for prediction

estimation and inference

simple linear regression

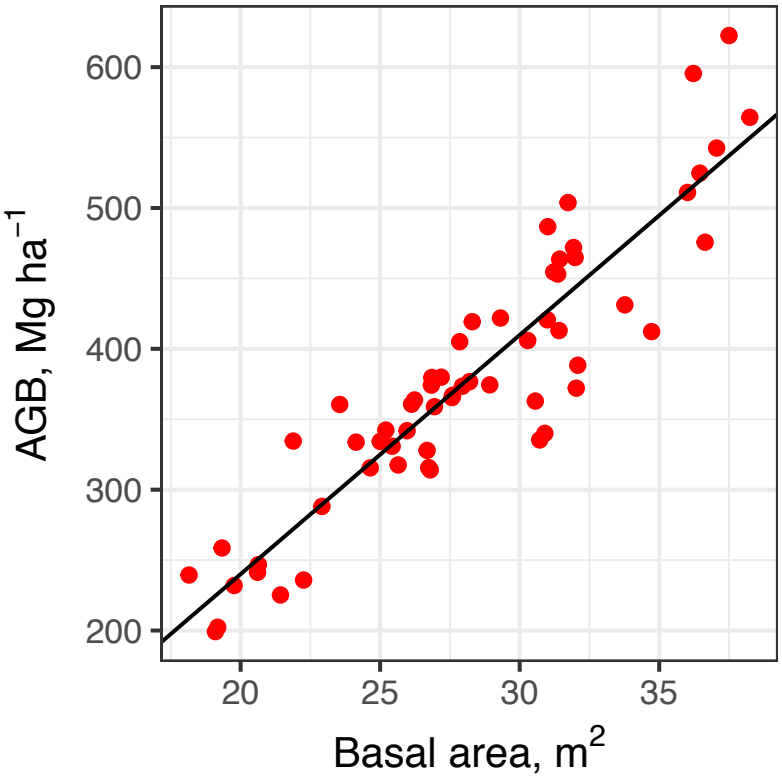
- back to biomass example
- simple linear regression used when response and explanatory variables are continuous

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

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

Greek letters

Y_i = AGB of i^{th} plot
X_i = basal area of i^{th} plot
α = population intercept
β = population slope
ε_i = residual variation



simple linear regression

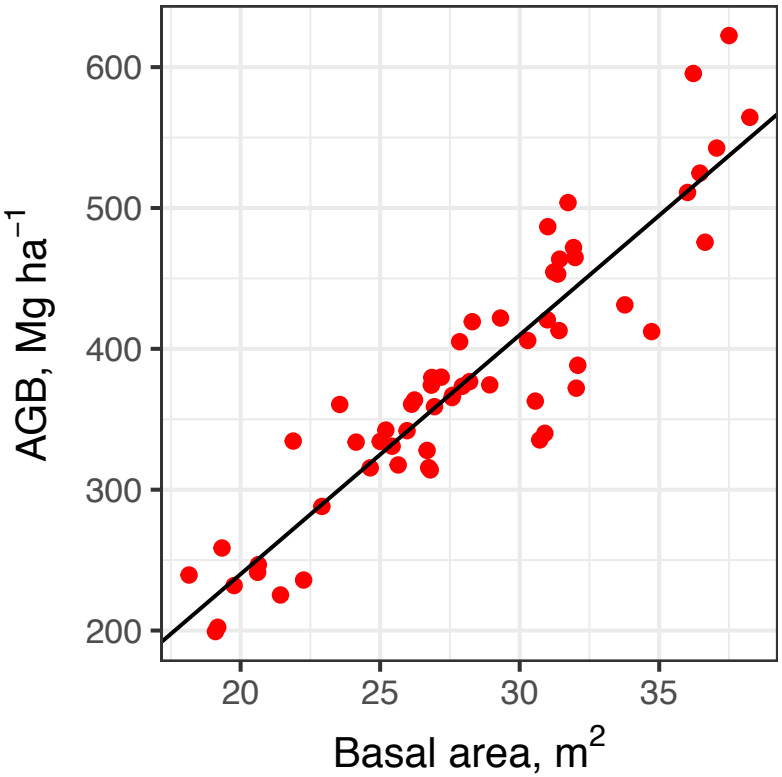
- back to biomass example
- simple linear regression used when response and explanatory variables are continuous

$$Y_i = a + bX_i + \varepsilon_i$$

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

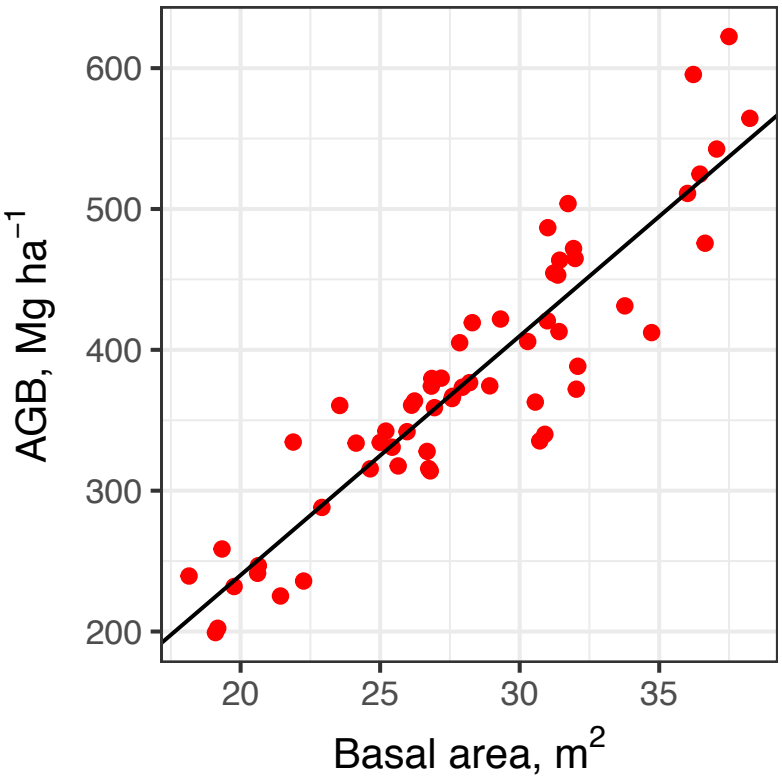
Roman letters

Y_i = AGB of i^{th} plot
X_i = basal area of i^{th} plot
a = sample intercept
b = sample slope
ε_i = residual variation



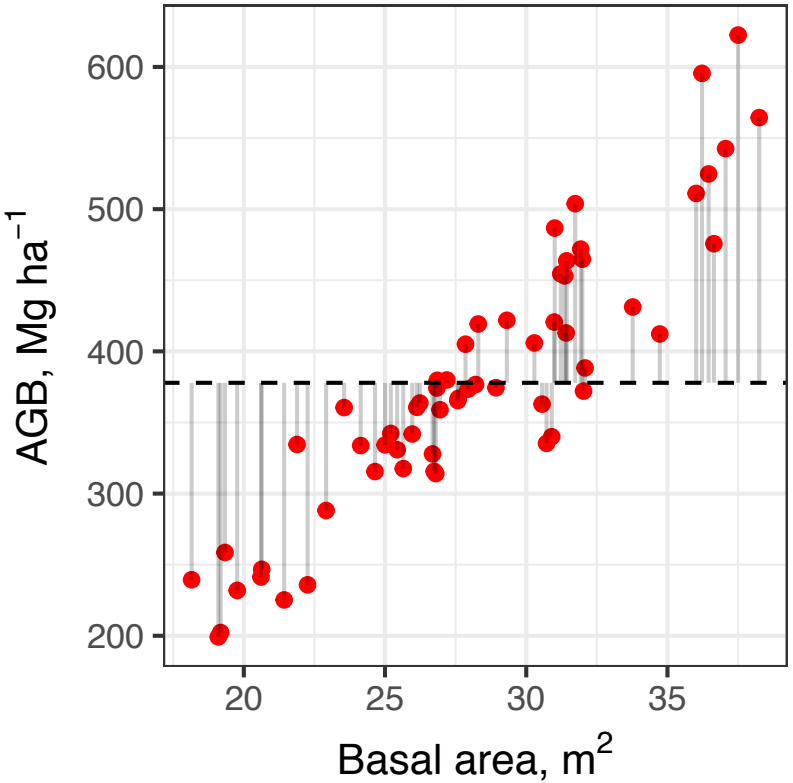
parameter estimation

- when we fitted the line of best fit by eye, we tried to minimize the distance between the line and the data points
- we can use maximum likelihood to estimate the parameters
- ML: given the data, what are the estimates of the intercept and slope that make the data most likely?



parameter estimation

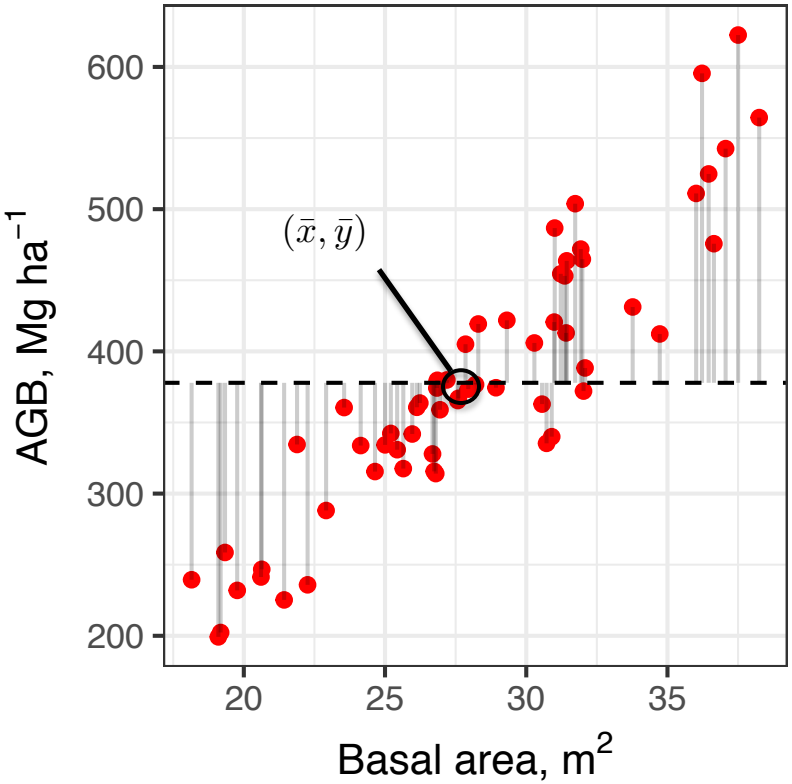
- with simple models, and as long as the data meet certain assumptions, the ML estimates of the parameters are provided by the method of least squares (ordinary LS)
- method of least squares defines the line of best fit by minimizing the error sums of squares
- let's see what that means…



null model: mean AGB

parameter estimation

- scatter around the overall mean is the total variation in *Y*
- total deviation is the total vertical distance of each point from the mean
- SST, or total sums of squares

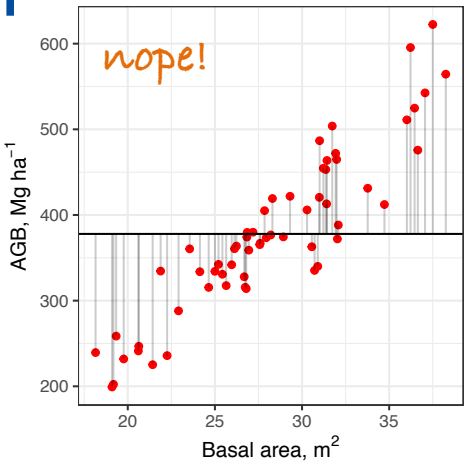


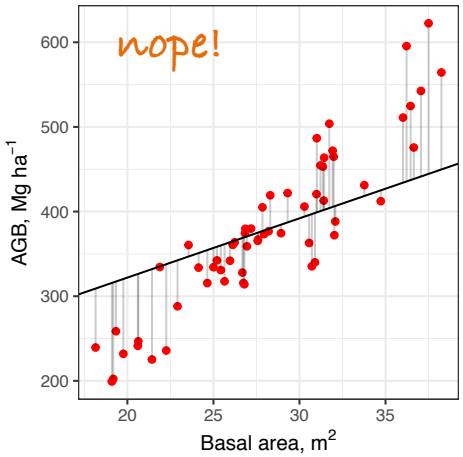
null model: mean AGB

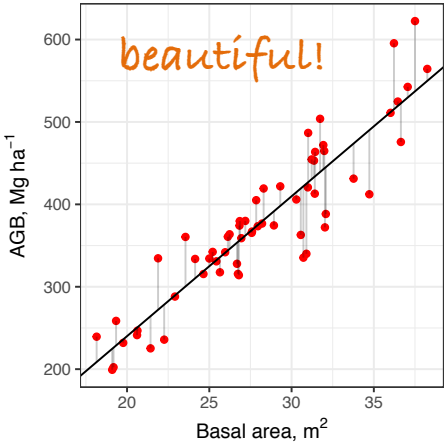
parameter estimation

- to find the line of best fit, we can rotate this line through the mean of *Y* and *X*
- the distance of the observed values to this line is SSE

  (SS error or residuals)
- we want to find the line that minimizes SSE

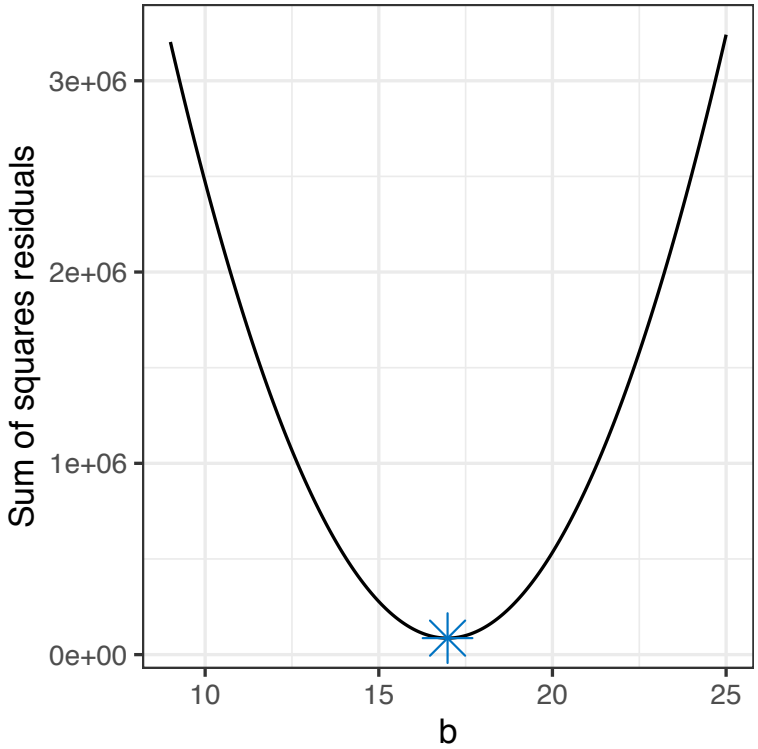






parameter estimation

- can plot the SSE for each value of *b*
- value of *b* that gives the lowest value of SSE is the ML estimate of *b*
- in practice, this is done using calculus, by finding the derivative of SSE with respect to *b*, set it to zero and solve for *b*
- ML estimate = 16.98
- estimates of *a* and *b* give the effect sizes



simple linear regression - inference

- now we assess whether *X* explains a significant proportion of the variability in *Y*
- test the null hypothesis that the slope of the line is equal to zero: (H_0: $\beta = 0$)
- ANOVA splits the total variance (SST) into the part explained by the regression (SSR) and the part not explained by *X* (SSE)
  - if the ratio of SSR/SSE is large, then reject H_0
  - if the ratio of SSR/SSE is small, then retain H_0

sum of squares

1. total variation in Y:

SST = observed – overall mean

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

SST can be split into:

2. variance explained by X

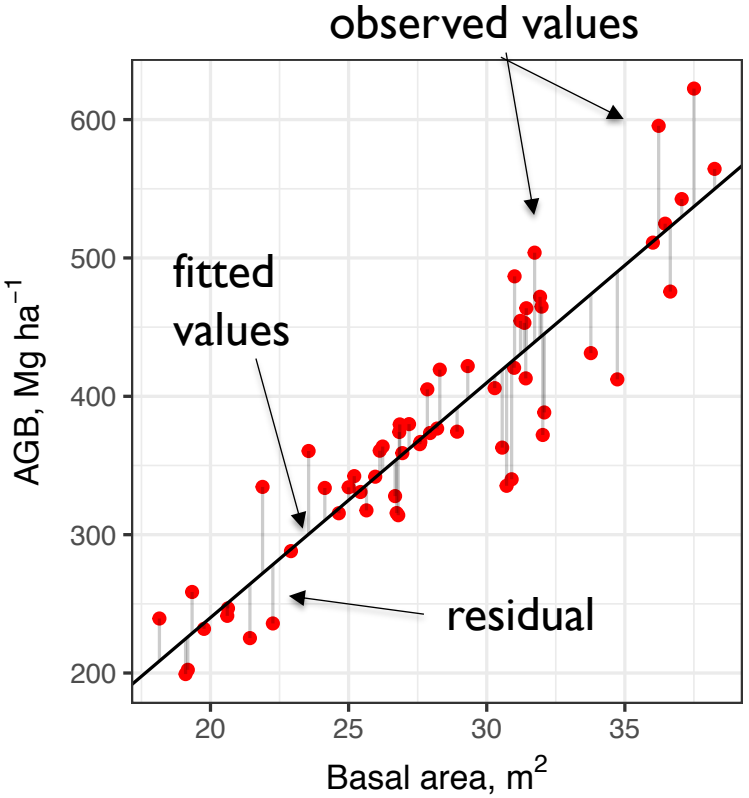
SSR = fitted – overall mean

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

3. variance not explained by X

SSE = observed - fitted

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



sum of squares

1. total variation in Y:

SST = observed – overall mean

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SST can be split into:

2. variance explained by X

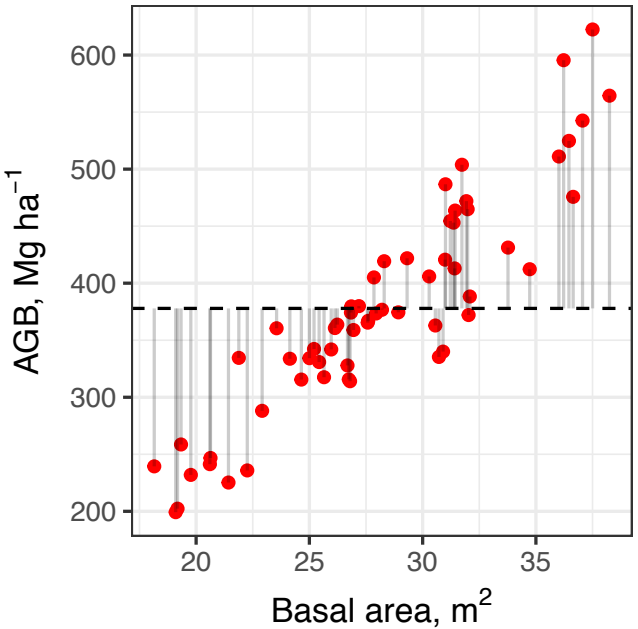
SSR = fitted – overall mean

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

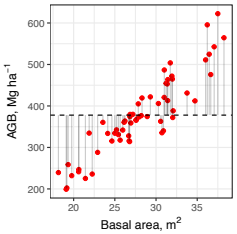
3. variance not explained by X

SSE = observed - fitted

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$



sum of squares



1 total variation in Y:

SST = observed – overall mean

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SST can be split into:

2 variance explained by X

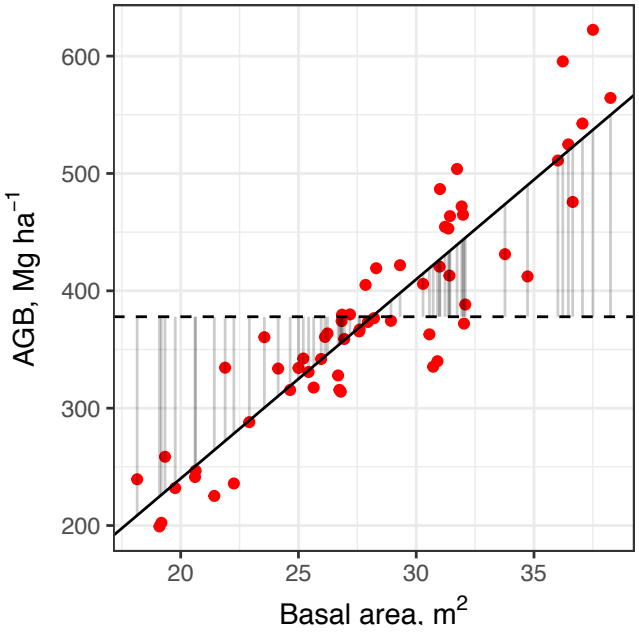
SSR = fitted – overall mean

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

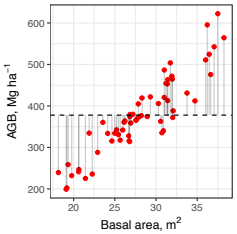
3 variance not explained by X

SSE = observed - fitted

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$



sum of squares

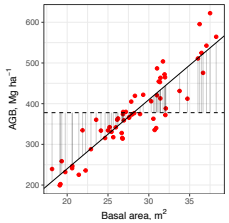


1 total variation in Y:

SST = observed – overall mean

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

SST can be split into:



2 variance explained by *X*

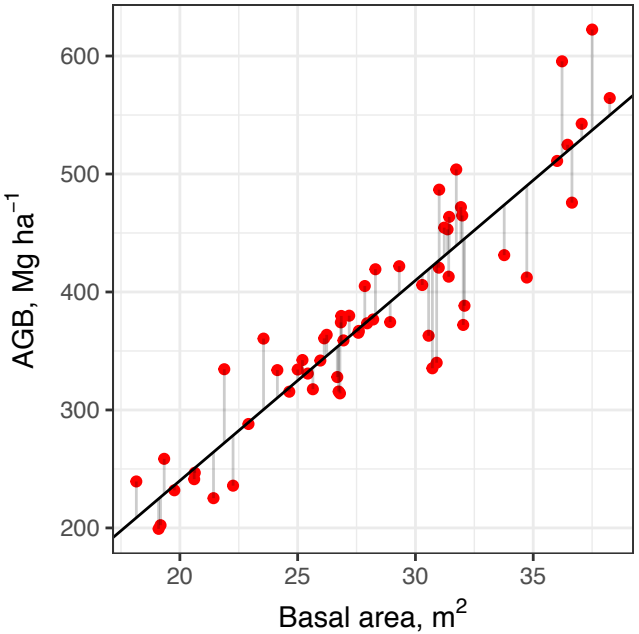
SSR = fitted – overall mean

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

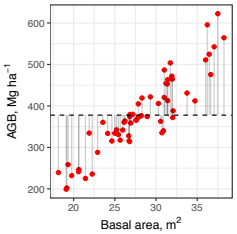
3 variance not explained by *X*

SSE = observed - fitted

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



sum of squares

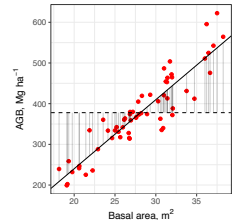


1 total variation in Y:

SST = observed – overall mean

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

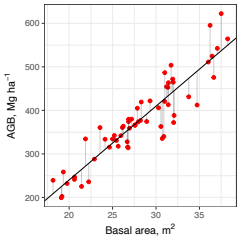
SST can be split into:



2 variance explained by *X*

SSR = fitted – overall mean

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$



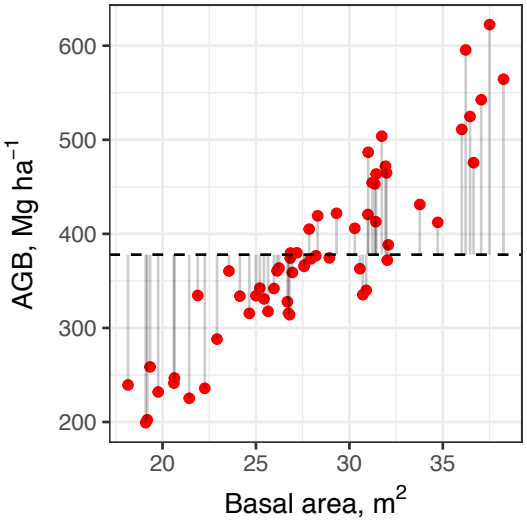
3 variance not explained by *X*

SSE = observed - fitted

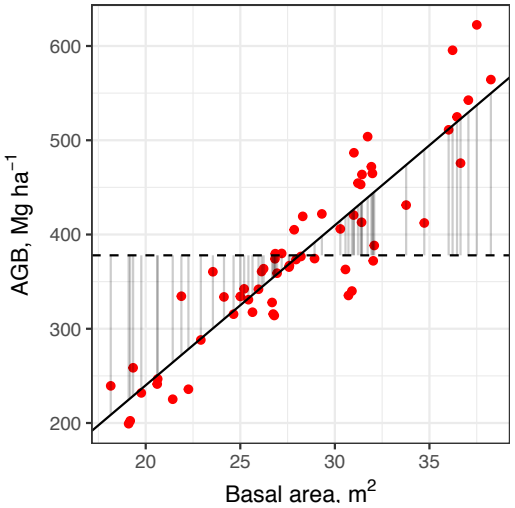
$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

SST = SSR + SSE

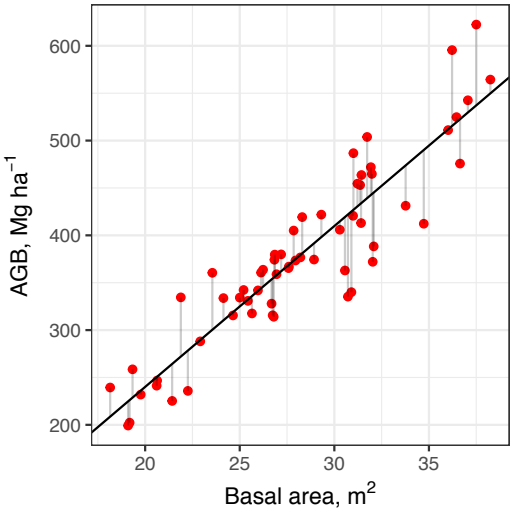
simple linear regression



SST
total sum of squares



SSR
regression sum of squares



SSE
error sum of squares

linear modeling in R

```
lm.mod <- lm(AGB ~ BasalArea, data = adat)
summary(lm.mod)
```

```
Call:lm(formula = AGB ~ BasalArea, data = adat)

Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept) -99.4735    28.0232   -3.55 0.000773 ***
BasalArea    16.9757     0.9807   17.31  < 2e-16 ***

---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 38.55 on 58 degrees of freedom
Multiple R-squared:  0.8378,  Adjusted R-squared:  0.835
F-statistic: 299.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

what matters here?

- parameter estimates of the intercept and slope
- fit of the model to the data
- coefficient of determination

```
Coefficients:

            Estimate  Std. Error t value   Pr(>|t|)
Intercept) -99.4735    28.0232   -3.55   0.000773 ***
BasalArea    16.9757     0.9807   17.31  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.55 on 58 degrees of freedom
Multiple R-squared:  0.8378, Adjusted R-squared:  0.835
F-statistic: 299.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

parameter estimates

- our parameter estimates (or coefficients):

```
Coefficients:

             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -99.4735    28.0232    -3.55     0.000773 ***
BasalArea    16.9757     0.9807    17.31     < 2e-16 ***
```

$$\mathrm{AGB} = -99.47 + 16.98 \cdot \mathrm{BA}$$

- for every 1-m^2 increase in basal area, AGB increases by 16.98 Mg ha^{-1}
- slope represents the change in *Y* for a one unit change in *X*

parameter estimates and inference

- our parameter estimates (or coefficients):

```
Coefficients:

             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -99.4735    28.0232    -3.55     0.000773 ***
BasalArea    16.9757     0.9807    17.31     < 2e-16 ***
```

- standard errors given in same units as coefficients
- t-value is the coefficient estimate / standard error
- test the null hypothesis that coefficient = 0
- p-value calculated with t-distribution
- reject null hypothesis with large t-value and small p-value

model fit to the data

- testing the null hypothesis (H_0: β=0) with ANOVA

```
Coefficients:

            Estimate  Std. Error t value   Pr(>|t|)
Intercept) -99.4735    28.0232   -3.55   0.000773 ***
BasalArea    16.9757     0.9807   17.31  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.55 on 58 degrees of freedom
Multiple R-squared:  0.8378, Adjusted R-squared:  0.835
F-statistic: 299.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

model fit to the data

- testing the null hypothesis (H_0: β=0) with ANOVA

```
> summary.aov(lm.mod)
```

```
             Df    Sum Sq   Mean Sq    F value   Pr(>F)
BasalArea     1    445367    445367     299.6   <2e-16 ***
Residuals    58    86206      1486
```

$$df_{reg} = 1$$
$$df_{error} = n - 2$$
$$df_{total} = n - 1$$

model fit to the data

- testing the null hypothesis (H_0: β=0) with ANOVA

```
> summary.aov(lm.mod)
```

SSR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BasalArea	1	445367	445367	299.6	<2e-16 ***
Residuals	58	86206	1486		

SSE

model fit to the data

- testing the null hypothesis (H_0: β=0) with ANOVA

```
> summary.aov(lm.mod)
```

SSR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BasalArea	1	445367	445367	299.6	<2e-16 ***
Residuals	58	86206	1486		

SSE

$$MS_{reg} = SSR/df_{reg} = 445367/1 = 445367$$
$$MS_{error} = SSE/df_{error} = 86206/58 = 1486$$

model fit to the data

- testing the null hypothesis (H_0: β=0) with ANOVA

```
> summary.aov(lm.mod)
```

SSR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BasalArea	1	445367	445367	299.6	<2e-16 ***
Residuals	58	86206	1486		

SSE

$$F = MS_{reg}/MS_{error} = 445367/1486 = 299.6$$

- F value assumed to follow F distribution with df_{reg} and df_{error}
- large F value, therefore, H_0: $\beta = 0$ is very unlikely

coefficient of determination – R^2

- fraction of total variance explained by the regression

$R^2 = SSR/SST$

```
Residual standard error: 38.55 on 58 degrees of freedom
Multiple R-squared:  0.8378, Adjusted R-squared:  0.835
F-statistic: 299.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

- therefore, basal area explains ~84% of the variation in AGB
- the more explanatory variables, the higher the value of R^2

LET'S RECAP...

1. estimate the regression parameters, slope and y-intercept

2. test null hypothesis of no effect of X on Y
   - partition sources of variance (i.e. ANOVA)
   - SST (total variance), SSR (regression sum of squares), and SSE (error residual sum of squares)
   - calculate F-statistic

3. calculate coefficient of determination, R^2

linear regression assumptions

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

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

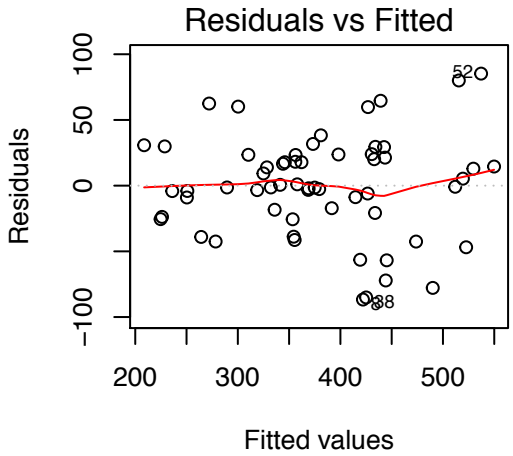
- residuals are normally distributed
- residuals are independent of each other
- constant variability: variances are constant along the regression line (homogeneity)
- explanatory variable X is measured without error

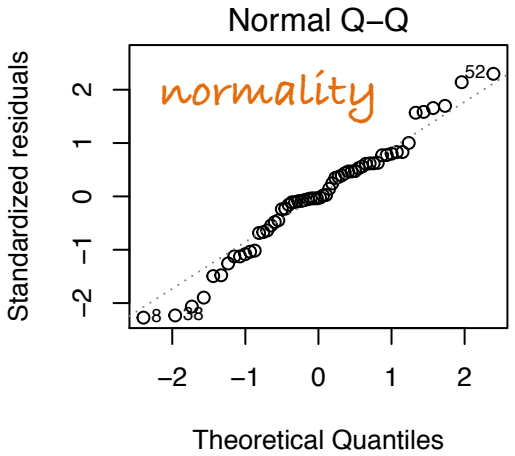
these assumptions must be checked before you accept that your model is valid

https://gallery.shinyapps.io/slr_diag/

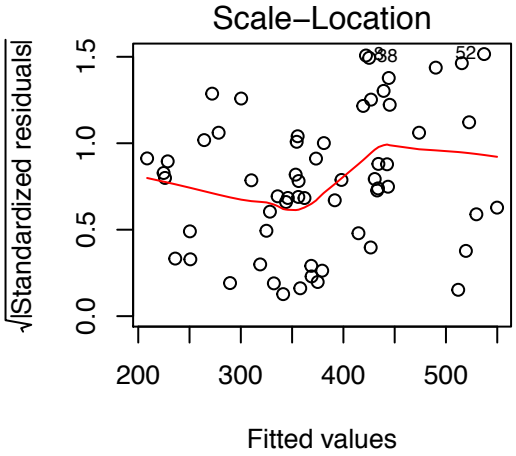
model validation in R

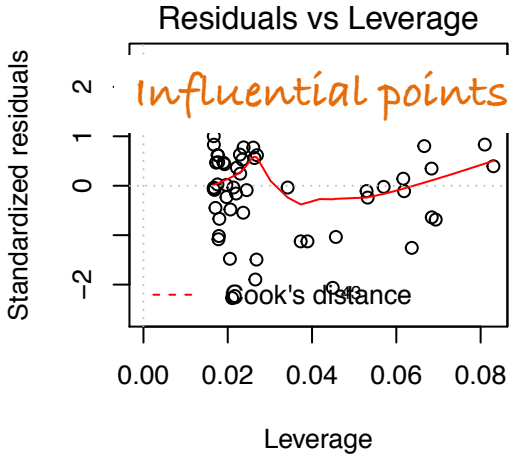
`plot(lm.mod)`



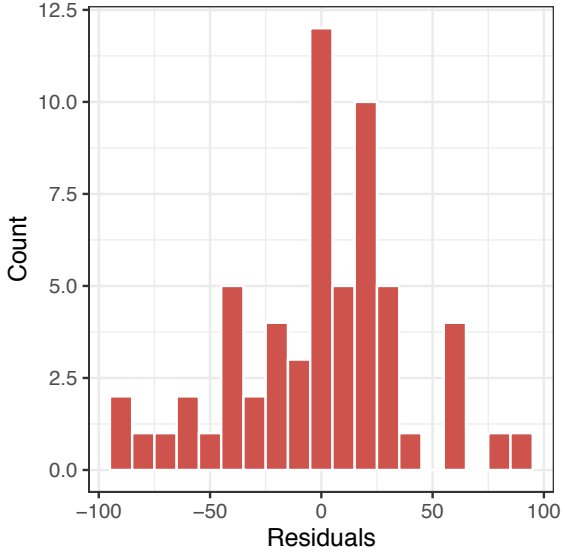


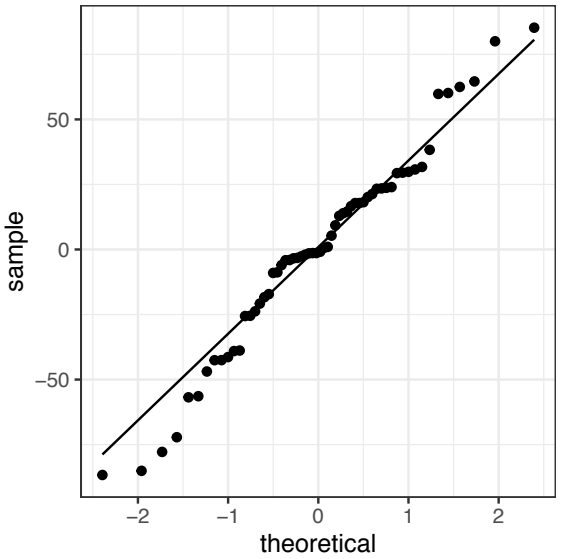
homogeneity

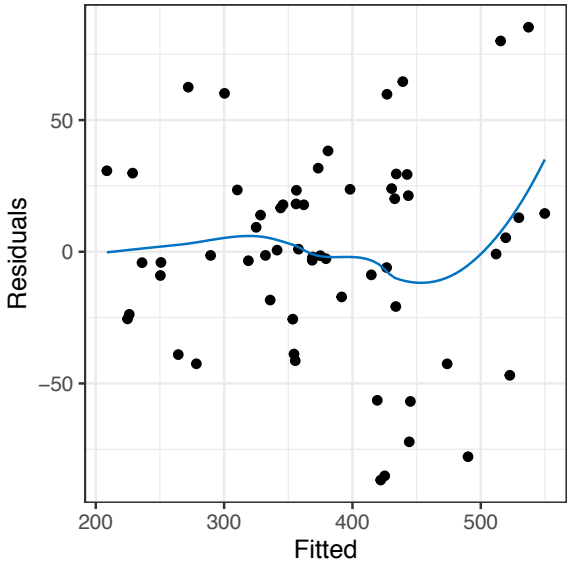




model validation in R







Questions?