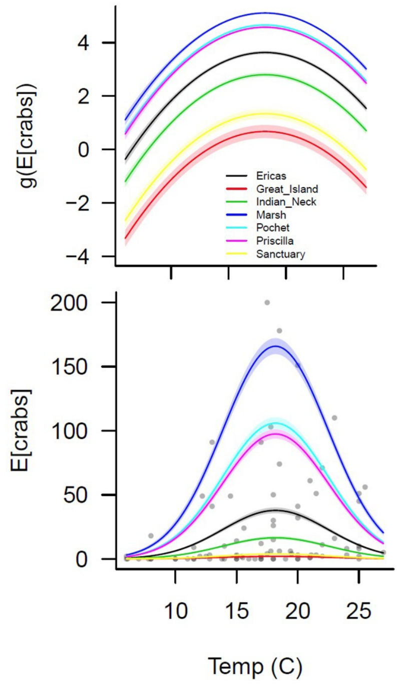
ECO 636
Applied Ecological Statistics

Week 2 – Linear Model Intro



Meg Graham MacLean, PhD

Department of Environmental Conservation

mgmaclean@umass.edu

2021 - Spring

The Week

Tuesday

- Review of data exploration
- Modeling process!
- Basics of a linear model
  - Null model - tree example

Wednesday (Lab)

- Data exploration

Thursday (asynchronous)

- More linear models (two groups)

A Protocol for Data Exploration

- Formulate a biological/ecological hypothesis & collect data
- Data Exploration
  1. Outliers (Y & X) — boxplot & Cleveland dotplot
  2. Homogeneity (Y) — conditional boxplot
  3. Normality of errors (Y) — histogram or QQ-plot
  4. Zero trouble (Y)*
  5. Collinearity (X) — scatterplots, correlations, VIF*
  6. Relationships (Y & X) — pair plots
  7. Interactions (X) — coplots
  8. Independence (Y)*
- Apply statistical model

*we will chat about these more later

Modeling process

We will adopt a formulaic approach to modeling

- Core concepts are similar for all models
- Repetition should help reinforce ideas
- HOWEVER: every analysis is unique and requires considerable thought

Modeling process

1. State the question/hypothesis
   - What is the question?
   - What are the variables (response and explanatory)?
2. Data exploration
   - Last class!
3. Describe the model
   - In word form (should come from your question)
   - In mathematical form
   - Identify the assumptions of the model
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

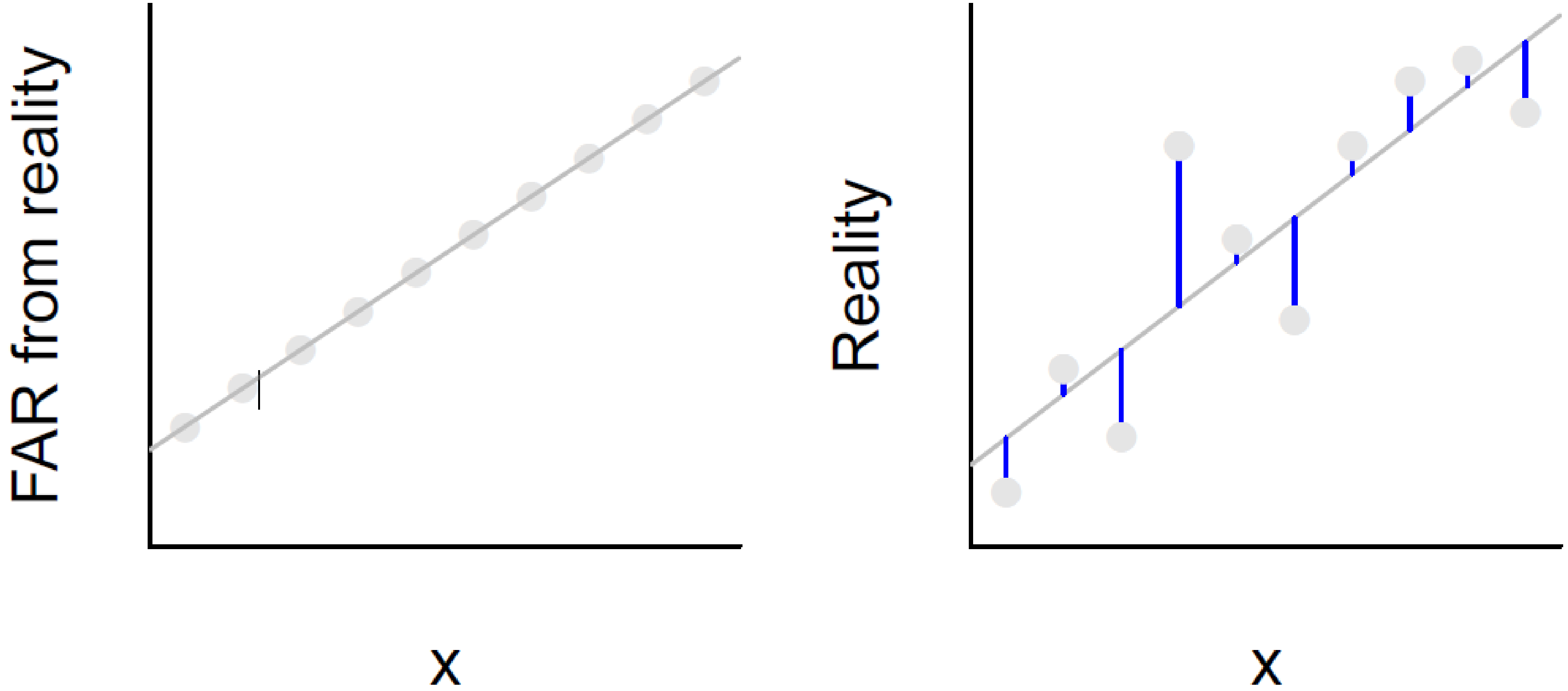
Let's assume we have done this already!

Modeling process

1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
   - *Last class!*
3. **Describe the model**
   - **In word form (should come from your question)**
   - **In mathematical form**
   - **Identify the assumptions of the model**
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

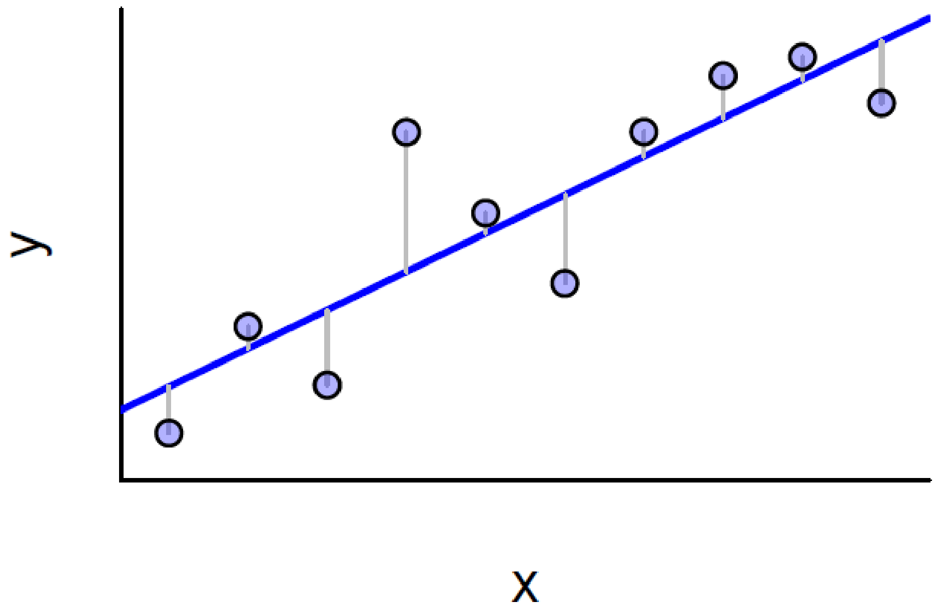
3. Describe the model

"If there was no variation, there would be no need for statistics" – Snee (1999)



3. Describe the model

- Linear models are the basis for many analytical methods
  - Two fundamental components:
    - Deterministic (signal) – the "expected" value of the response given X
    - Stochastic (noise) – the difference between the "observed" value of the response and the expected

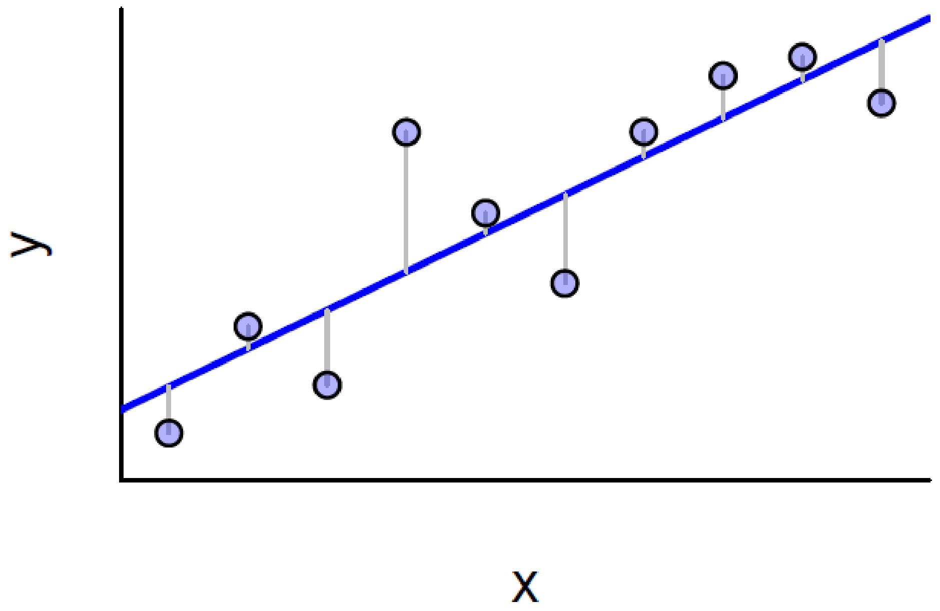


To find the best models we need statistics!

Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

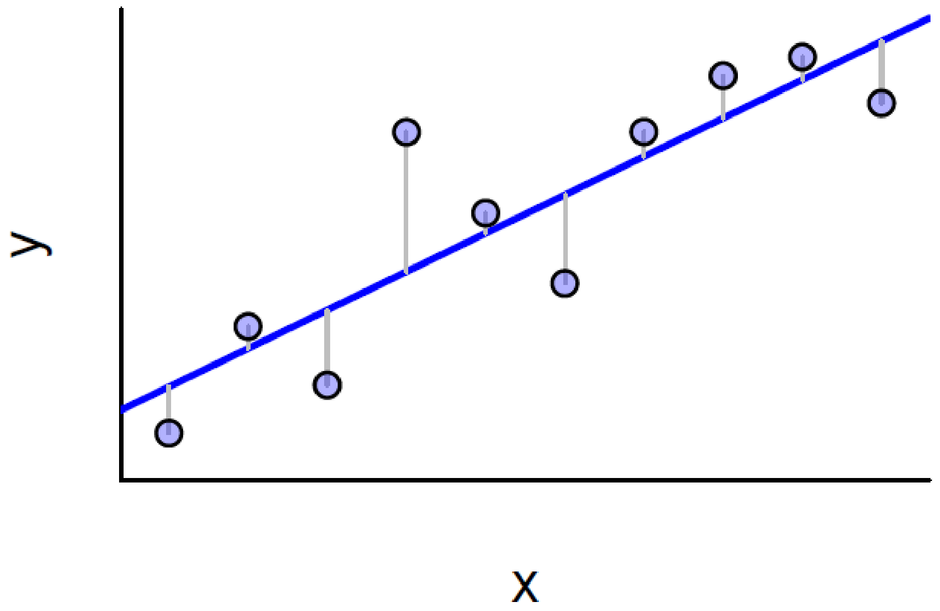
- Two fundamental components:
  - Deterministic (signal) – the "expected" value of the response given X
  - Stochastic (noise) – the difference between the "observed" value of the response and the expected



Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

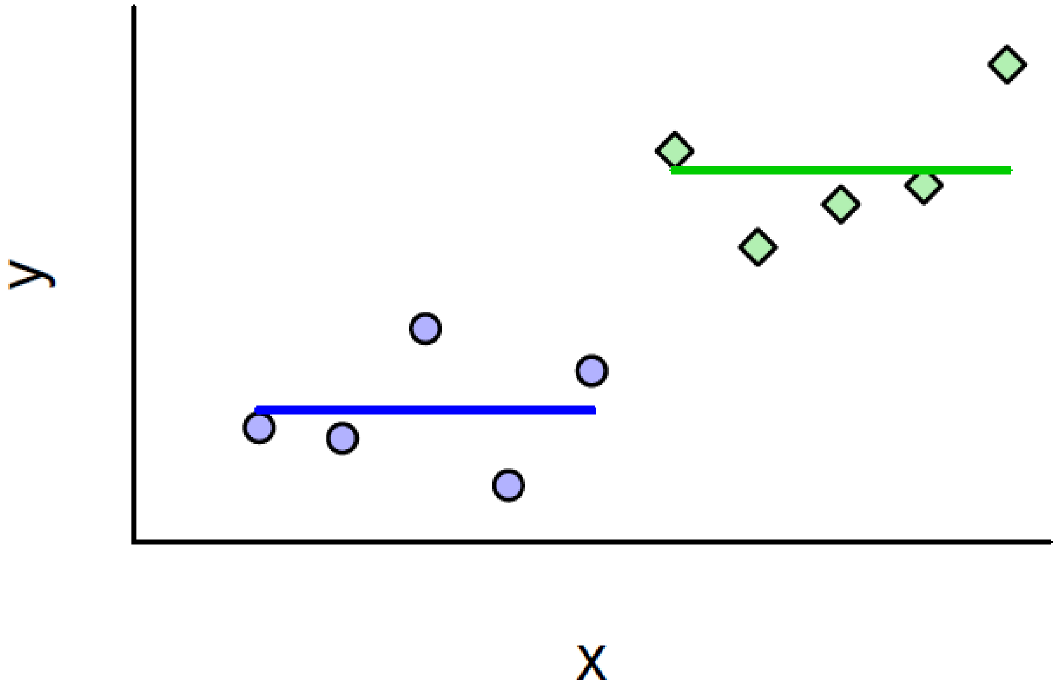
- Deterministic
  - β_0 and β_1 are parameters to be estimated
  - When X is *continuous* (mean = ———)



Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

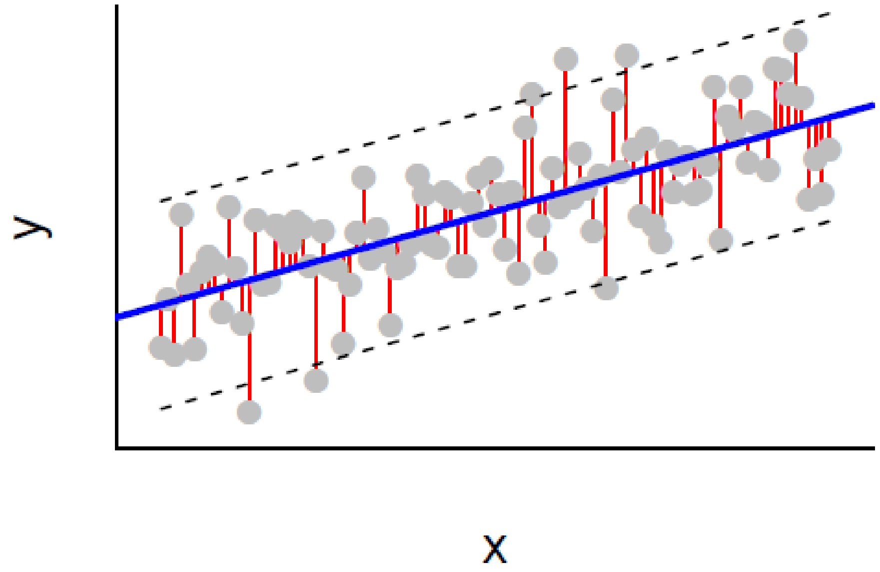
- Deterministic
  - β_0 and β_1 are parameters to be estimated
  - When X is *factor* (means = —— ——)

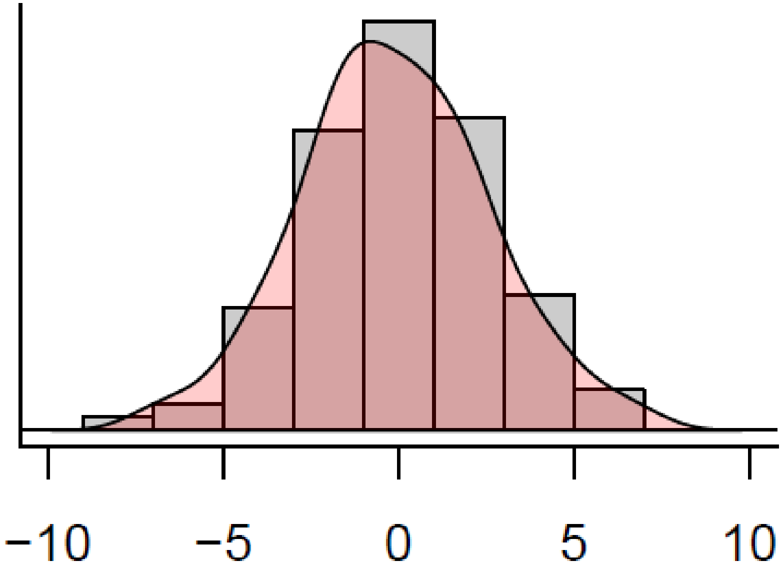


Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$
$$e_i = y_i - \hat{y}_i$$

- Stochastic – usually called the *residual*
  - Usually assume residuals are normally distributed (i.e., $N(0, \sigma^2)$)

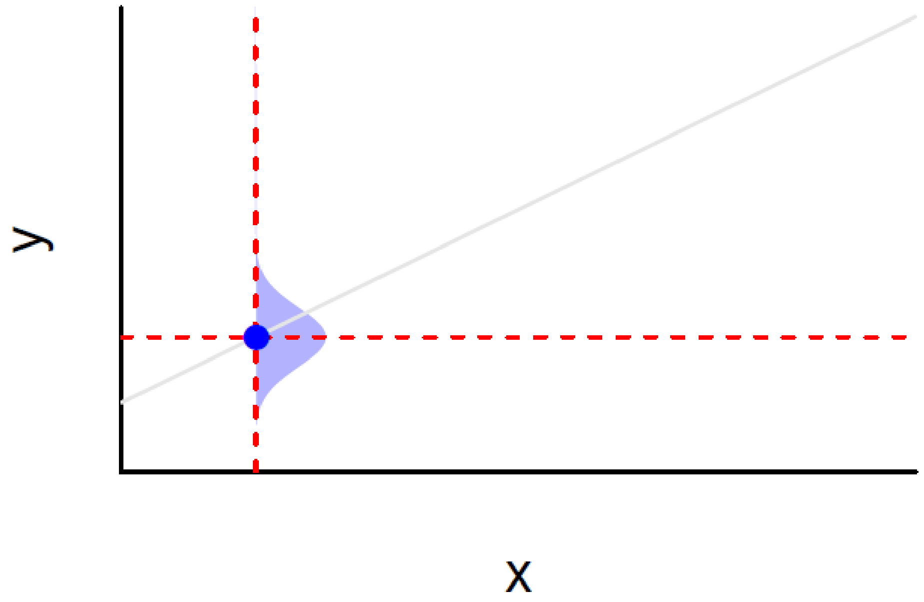




Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$
$$e_i = y_i - \hat{y}_i$$

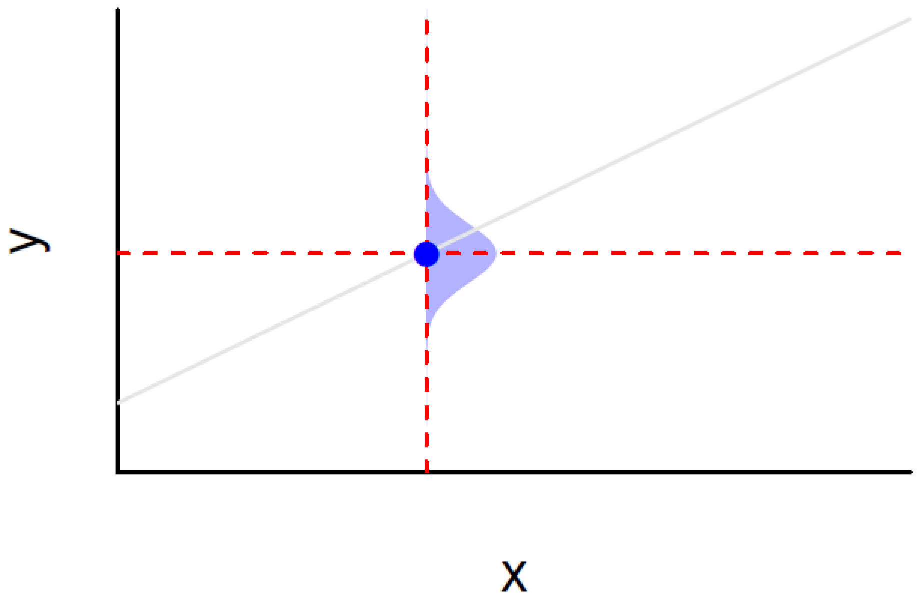
- Stochastic – usually called the *residual*
  - Usually assume residuals are normally distributed (i.e., $N(0, \sigma^2)$)



Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$
$$e_i = y_i - \hat{y}_i$$

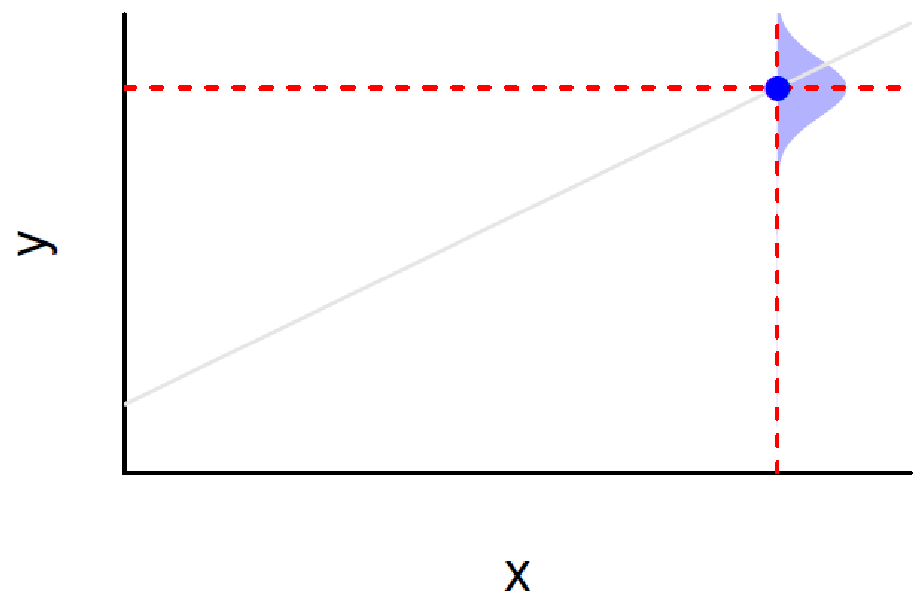
- Stochastic – usually called the *residual*
  - Usually assume residuals are normally distributed (i.e., $N(0, \sigma^2)$)



Simple linear model

$$y_i = \beta_0 + \beta_1 X_i + e_i$$
$$e_i = y_i - \hat{y}_i$$

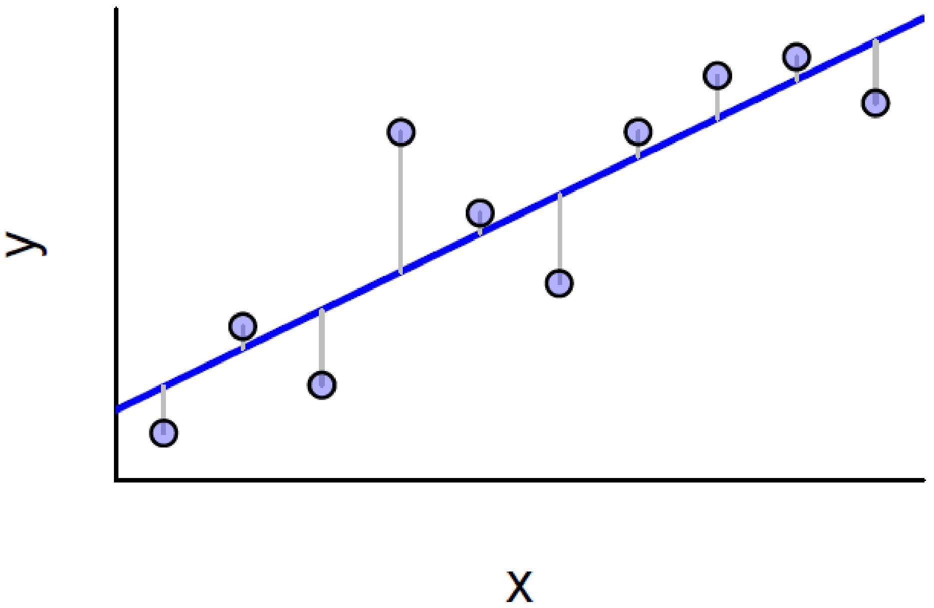
- Stochastic – usually called the *residual*
  - Usually assume residuals are normally distributed (i.e., $N(0, \sigma^2)$)



Simple linear model

Does a linear model have to be linear?

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

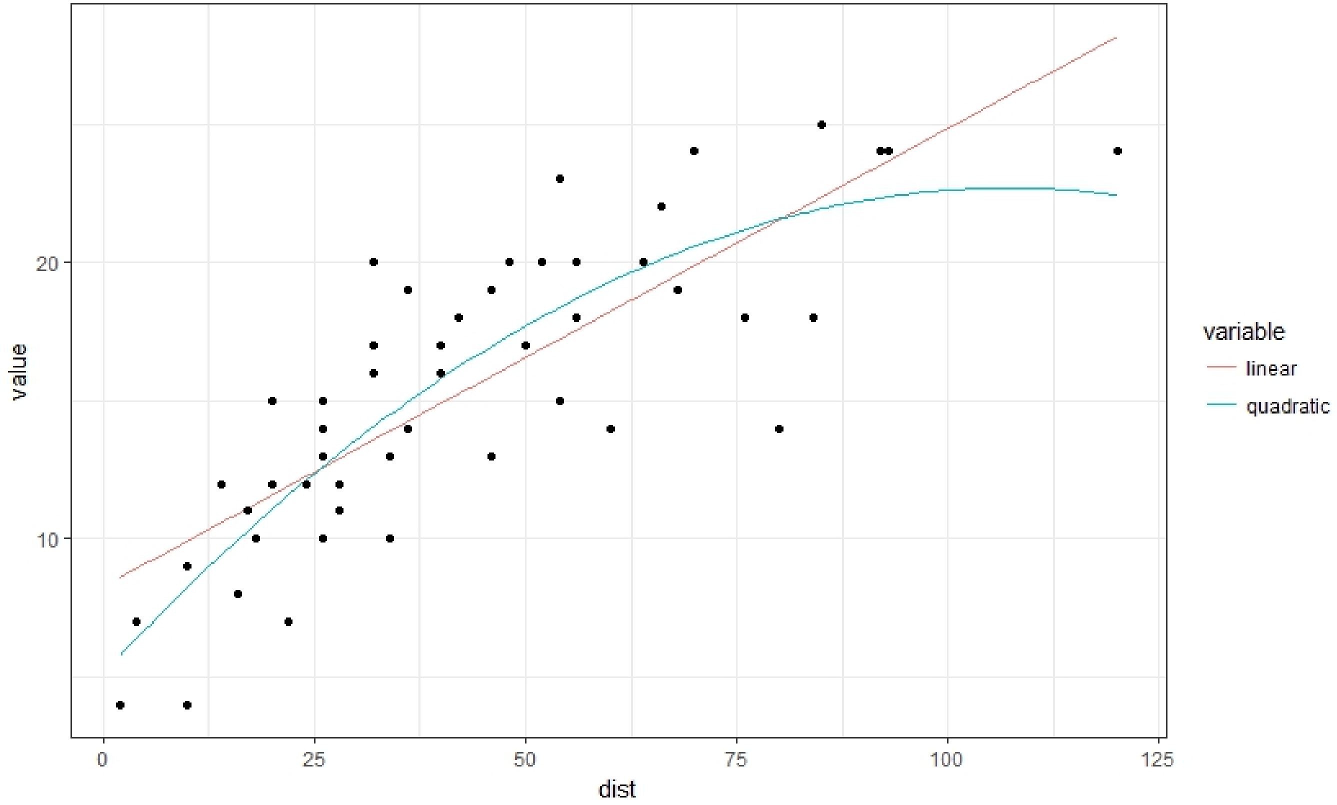


Simple linear model

Does a linear model have to be linear? No ☺

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i1}^2 + e_i$$

- Y just needs to be expressed as a linear function of X, but that function can be curvy



Assumptions

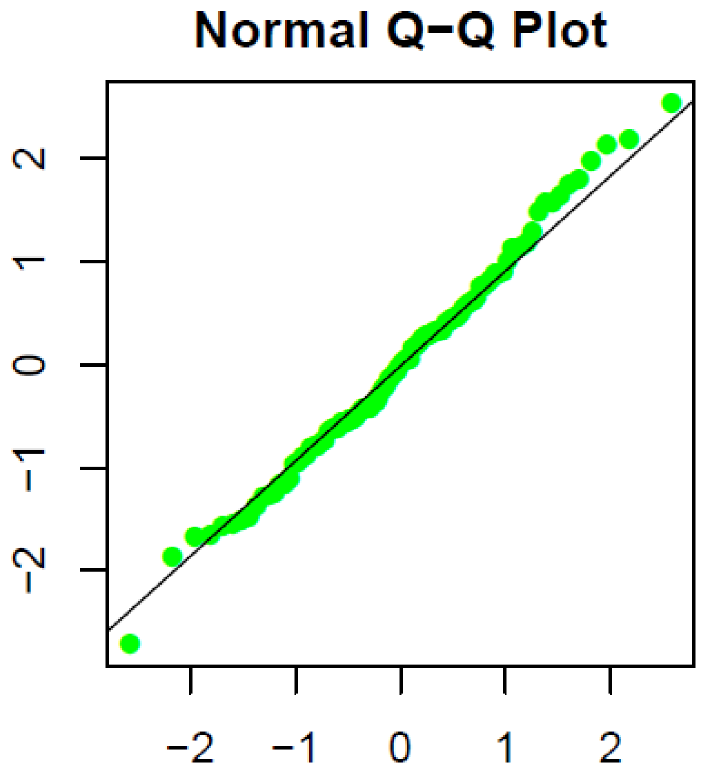
Basic assumptions of a normal linear model:

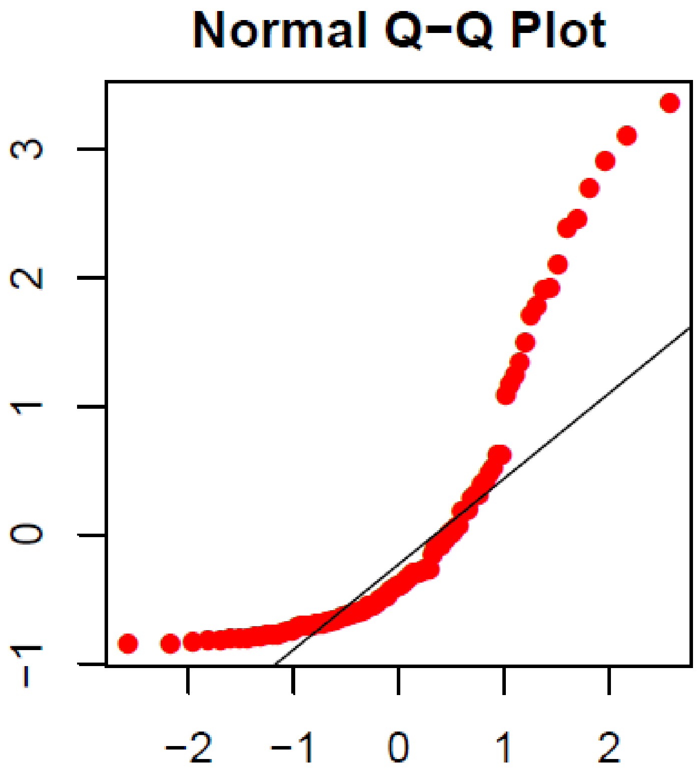
- Residuals are normally distributed
- Constant variance in residuals (homogeneity)
- Observations are independent
- Predictors are measured without error (fixed X)

Assumptions

Basic assumptions of a normal linear model:

- Residuals are normally distributed
  - Should match standard normal distribution

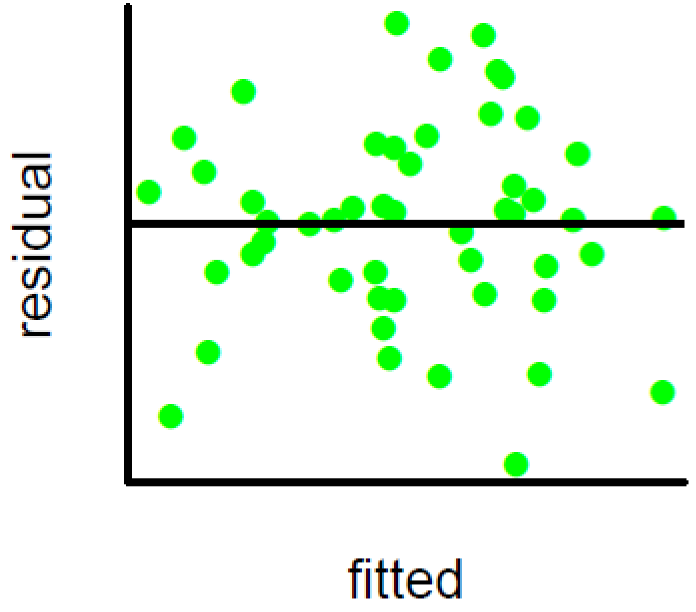


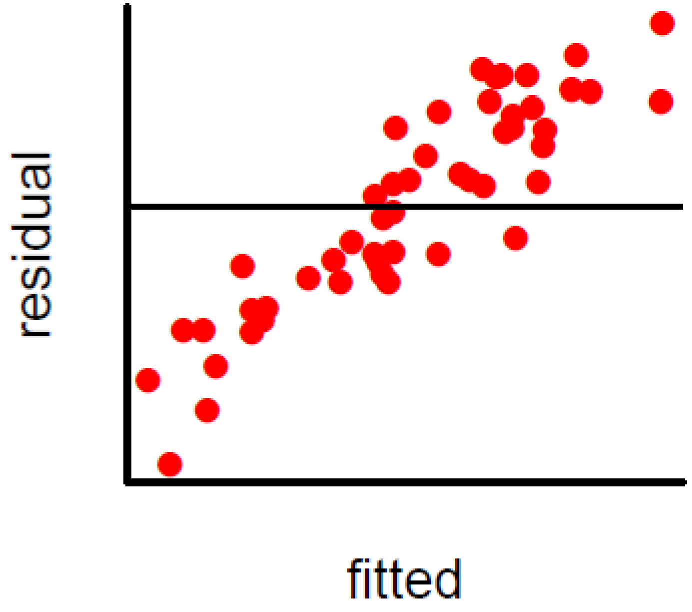


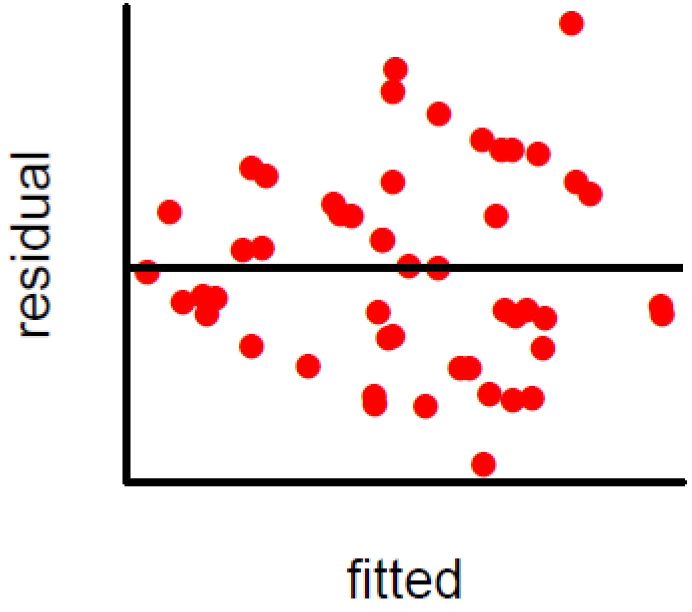
Assumptions

Basic assumptions of a normal linear model:

- Residuals are normally distributed
  - Should match standard normal distribution
- Constant variance in residuals (homogeneity)
  - Random scatter of points, no shape when plotting residuals vs. estimates/fitted points







Assumptions

Basic assumptions of a normal linear model:

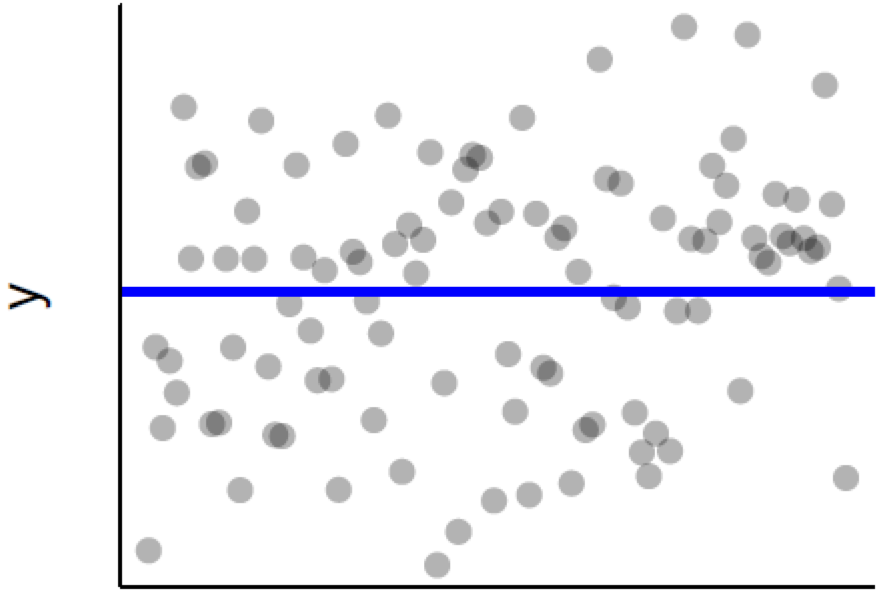
- Residuals are normally distributed
  - Should match standard normal distribution
- Constant variance in residuals (homogeneity)
  - Random scatter of points, no shape when plotting residuals vs. estimates/fitted points
- Observations are independent
  - No pseudo-replication, spatial/temporal autocorrelation
- Predictors are measured without error (fixed X)
  - Avoided through training and experimental design

Null linear model

The most basic linear model is the null model:

- Model of the mean – no explanatory variable of interest
- Single parameter special case of the linear model – intercept only
- What can we use a linear model to estimate?
  - Mean of the response variable
  - Variance of the response variable

$$y_i = \beta_0 + e_i$$
$$e_i \sim N(0, \sigma^2)$$



Let's try it – tree heights

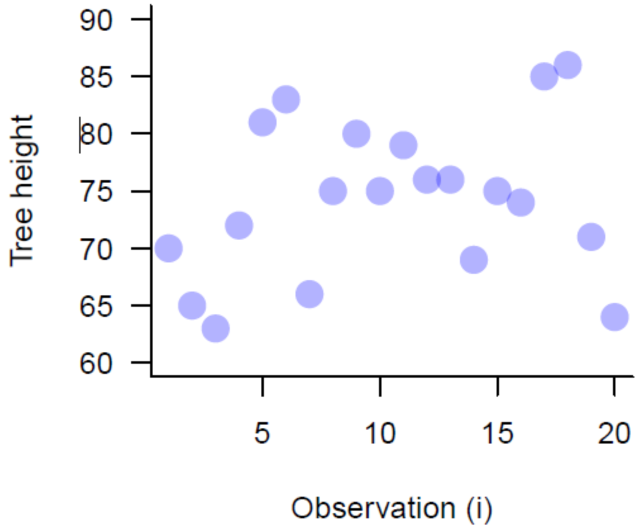
Let's say we go into a forest stand that is of interest to us (perhaps we want to harvest some wood). We measure the height of 20 randomly selected trees.

1. State the question/hypothesis
   - What is the expected height of a tree in the stand?
   - Variable: tree height (response)

By hand:

$$\bar{y} = \sum_i \frac{1}{n} y_i = 74.25$$

$$\sigma_y^2 = \frac{1}{n-1} \sum_i (y - \bar{y})^2$$

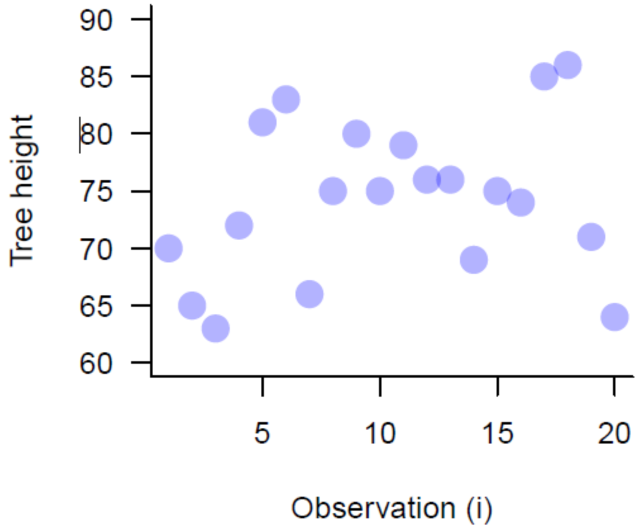


Let's try it – tree heights

Let's say we go into a forest stand that is of interest to us (perhaps we want to harvest some wood). We measure the height of 20 randomly selected trees.

1. State the question/hypothesis
   - What is the expected height of a tree in the stand?
   - Variable: tree height (response)

But let's use a linear model!



Modeling process

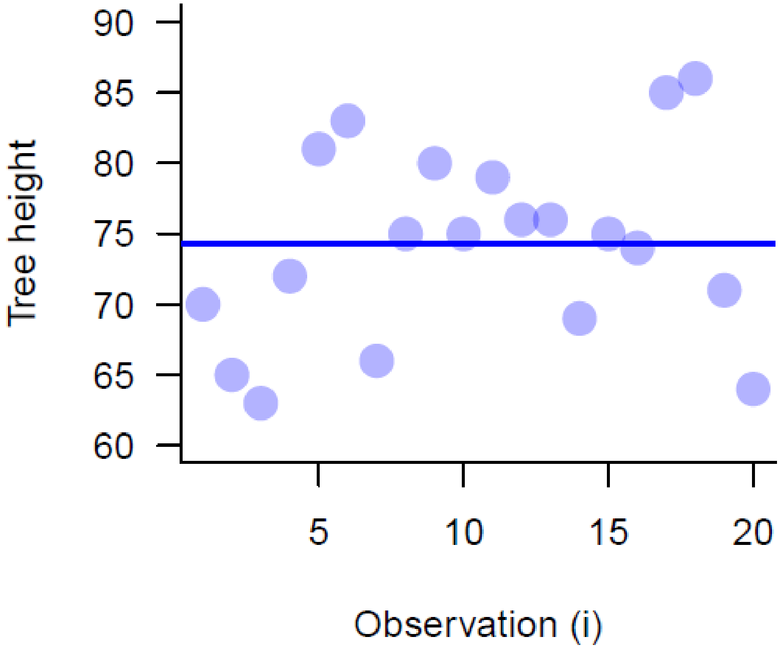
1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
   - *Last class!*
3. Describe the model
   - In word form (should come from your question)
   - In mathematical form
   - Identify the assumptions of the model
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
   - Model validation
   - Model selection
6. Interpret the results

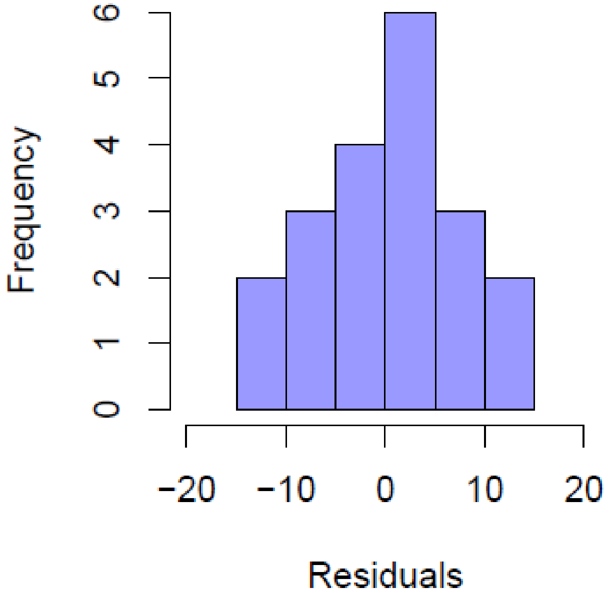
3. Describe the model

1. Describe the model – in word form:
   - What is the expected height of a randomly selected tree?
2. Describe the model – in mathematical form:

$$y_i = \beta_0 + e_i$$

- y_i is height (response)
- β_0 is the intercept
- e_i is the residuals





3. Describe the model

1. Describe the model – in word form:
   - What is the expected height of a randomly selected tree?
2. Describe the model – in mathematical form:

$$y_i = \beta_0 + e_i$$

   - y_i is height
   - β_0 is the intercept
   - e_i is the residuals
3. What are the assumptions?
   - Residuals are normally distributed
   - Constant variance (homogeneity)
   - Observations are independent
   - Predictors measured without error (fixed X)

Modeling process

1. *State the question/hypothesis*
    - *What is the question?*
    - *What are the variables (response and explanatory)?*
2. *Data exploration*
    - *Last class!*
3. *Describe the model*
    - *In word form (should come from your question)*
    - *In mathematical form*
    - *Identify the assumptions of the model*
4. Fit the model! (In R, of course ☺)
5. Evaluate the output
    - Model validation
    - Model selection
6. Interpret the results

4. Fit the model

Algebra: $y_i = \beta_0 + e_i$

R:

```
> m0 <- lm(height ~ 1, data = trees)
> m0

Call:
lm(formula = height ~ 1, data = trees)

Coefficients:
(Intercept)
      74.25
```

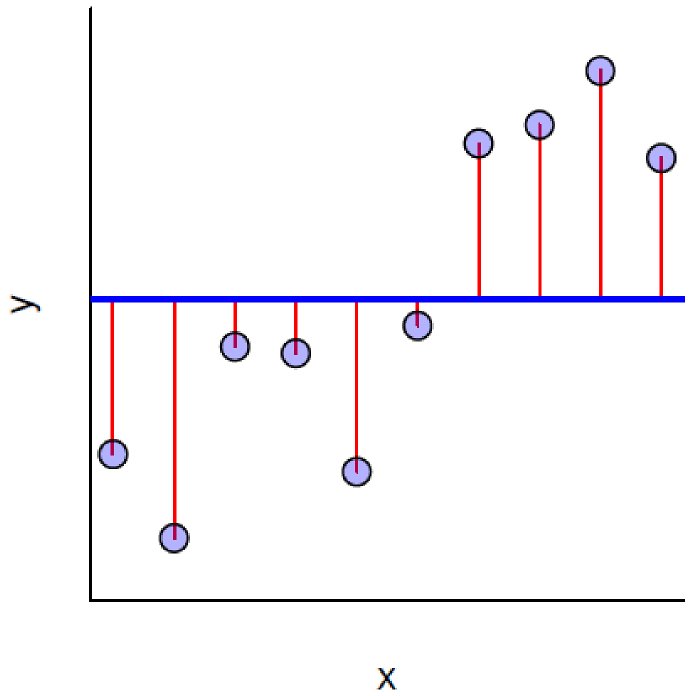
How do we estimate parameters (e.g., β_0)?

- Ordinary least squares (OLS) for the standard linear model
- Maximum likelihood (ML) or OLS for generalized linear models

Sum of Squares (refresh)

Sum of Squares (refresh)

- Method to characterize variability of the data
  - Total sum of squares: $SS_{total} = \sum(y_i - \bar{y})^2$

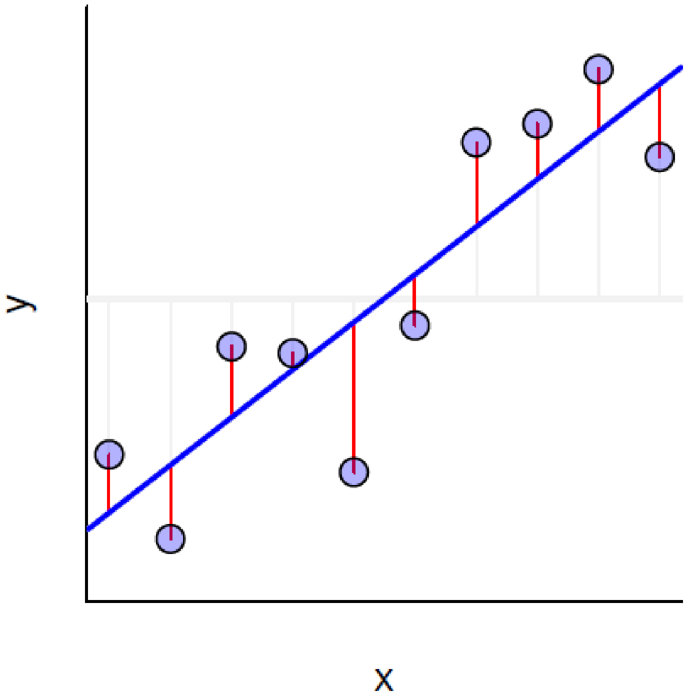


Sum of Squares (refresh)

Sum of Squares (refresh)

- Method to characterize variability of the data
  - Total sum of squares: $SS_{total} = \sum(y_i - \bar{y})^2$
  - Residual sum of squares: $SS_{residual} = \sum\left(y_i - x_i^T b\right)^2$

OLS minimizes $SS_{residual}$ to fit the model



4. Fit the model

In R, fitted model objects have some generic functions (to help with steps 5 and 6: evaluate the output and interpreting results):

- `plot()` produces diagnostic plots
- `summary()` produces a summary table
- `coef()` returns the estimated coefficients
- `fitted()` returns the fitted/predicted values
- `resid()` returns the residuals
- `predict()` returns predictions for a given set of covariates

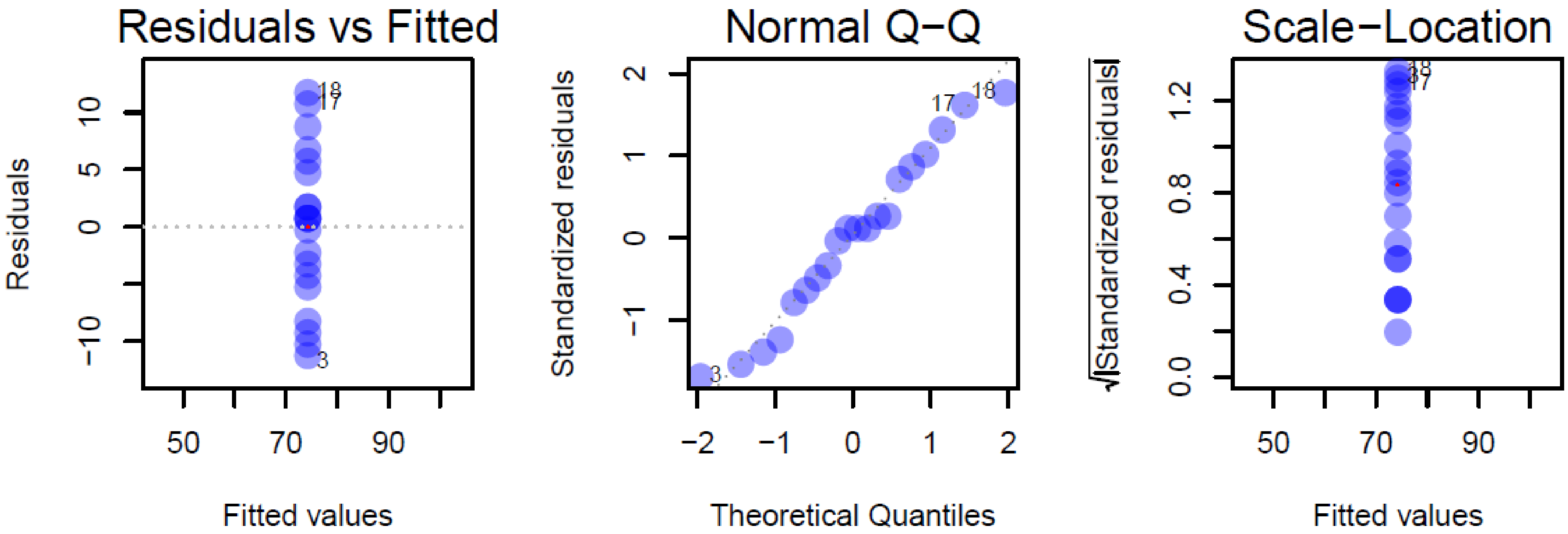
Modeling process

1. *State the question/hypothesis*
    - *What is the question?*
    - *What are the variables (response and explanatory)?*
2. *Data exploration*
    - *Last class!*
3. *Describe the model*
    - *In word form (should come from your question)*
    - *In mathematical form*
    - *Identify the assumptions of the model*
4. *Fit the model! (In R, of course ☺)*
5. Evaluate the output
    - Model validation
    - Model selection
6. Interpret the results

5. Evaluate the output

Model validation – what are the assumptions and are they met?

```
> plot(lm(height ~ 1, data = trees))
```



5&6. Evaluating output and Interpreting results

```
> summary(lm(height~1,data = trees))

Call:
lm(formula = height ~ 1, data = trees)

Residuals:
   Min      1Q  Median      3Q     Max 
-11.25  -4.50   0.75   5.00  11.75 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   74.250      1.527   48.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.828 on 19 degrees of freedom
```

Distribution of the residuals

5&6. Evaluating output and Interpreting results

```
> summary(lm(height~1,data = trees))

Call:
lm(formula = height ~ 1, data = trees)

Residuals:
   Min     1Q Median     3Q    Max 
-11.25  -4.50   0.75   5.00  11.75 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   74.250      1.527   48.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.828 on 19 degrees of freedom
```

Parameter estimate

Standard errors of the estimate

t-statistic to test if coefficient is significantly different from 0

p-value, or the probability of getting *t* large (or larger) if null hypothesis (β_0 = 0) is true

5&6. Evaluating output and Interpreting results

```
> summary(lm(height~1,data = trees))

  Call:
  lm(formula = height ~ 1, data = trees)

  Residuals:
     Min      1Q  Median      3Q     Max
  -11.25  -4.50   0.75    5.00   11.75

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)   74.250      1.527   48.63   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 6.828 on 19 degrees of freedom
```

$$\sigma = \frac{\sqrt{SS_{residuals}}}{df_{residuals}}$$

5&6. Evaluating output and Interpreting results

Sum of Squares (refresh)

- Method to characterize variability of the data
  - Total sum of squares: $SS_{total} = \sum(y_i - \bar{y})^2$
  - Residual sum of squares: $SS_{residual} = \sum\left(y_i - x_i^T b\right)^2$
  - Model sum of squares: $SS_{model} = SS_{total} - SS_{residual}$
    - The variation explained by the model

Model validation:

For a single predictor: $R^2 = \frac{SS_{model}}{SS_{total}}$

For multiple predictors: $AdjR^2 = \frac{SS_{model}/(n-(p+1))}{SS_{total}/(n-1)}$

n is sample size and p is number of explanatory variables

Modeling process

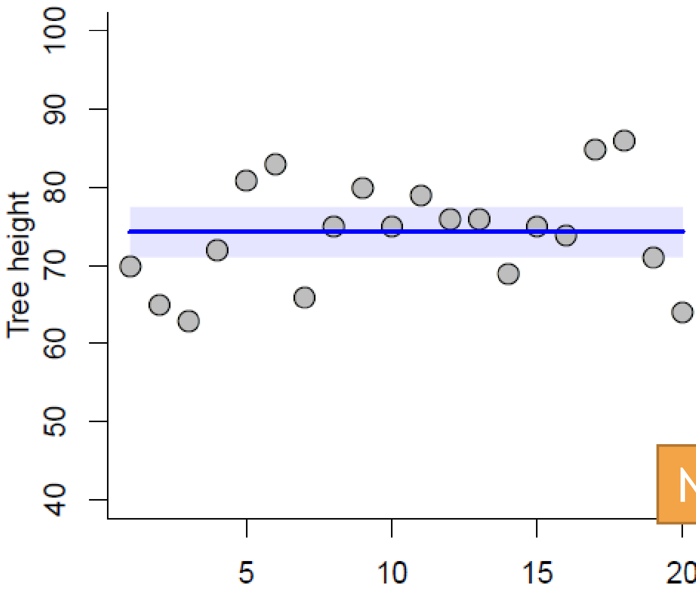
1. *State the question/hypothesis*
   - *What is the question?*
   - *What are the variables (response and explanatory)?*
2. *Data exploration*
   - *Last class!*
3. *Describe the model*
   - *In word form (should come from your question)*
   - *In mathematical form*
   - *Identify the assumptions of the model*
4. *Fit the model! (In R, of course ☺)*
5. *Evaluate the output*
   - *Model validation*
   - *Model selection*
6. Interpret the results

6. Interpreting results

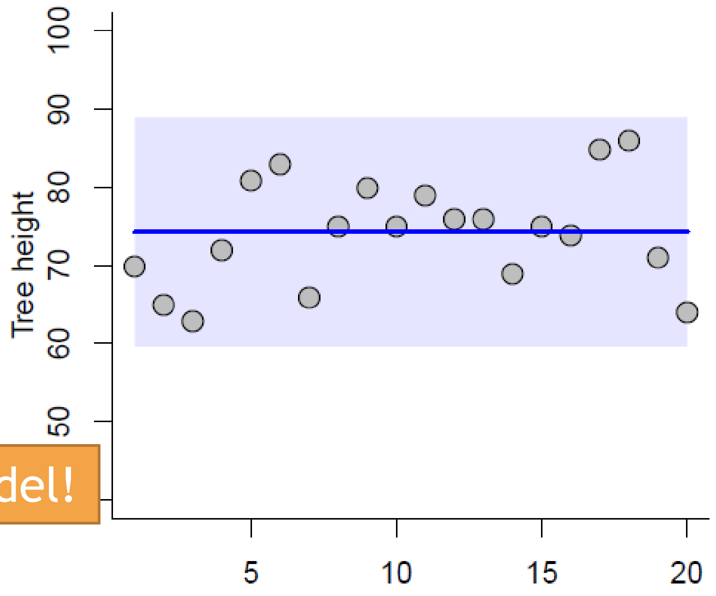
```
> predict(model,           #fitted model
+         newdata,         #newdata
+         se.fit=TRUE,     #calculate se's?
+         interval = c("none", "confidence", "prediction"),
+         level=0.95)
```

“confidence” = 95% CI around estimate

“prediction” = 95% CI around predictions



Null model!



Two samples!

Let's try this again, but with two groups (not the null model)

So far…

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`

Two samples!

Let's try this again, but with two groups (not the null model)

Next!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`

Two samples, where data collected is associated with membership in one of two groups (e.g., tall vs. short, stand 1 vs. stand 2)

Compare the population means = *t*-test as a linear model!

- H_0 = no difference between sample means
- H_1 = sample means differ

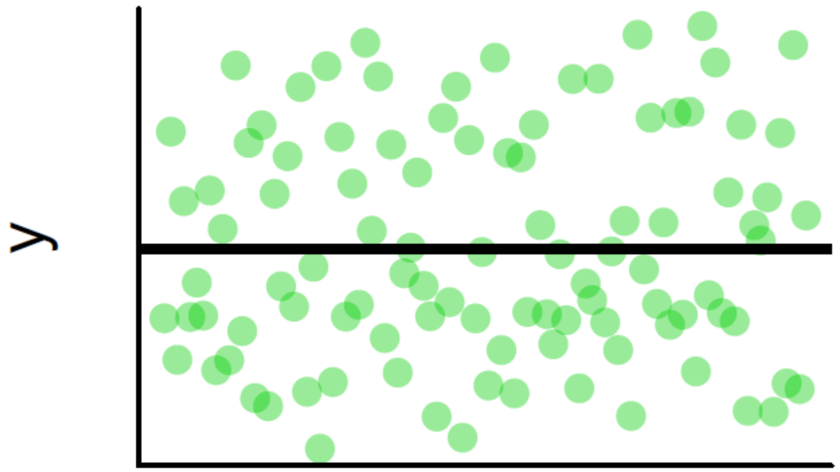
Review!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`

What does the first (null model) look like mathematically?

$$y_i = \beta_0 + e_i$$

What does the first (null model) look like graphically?



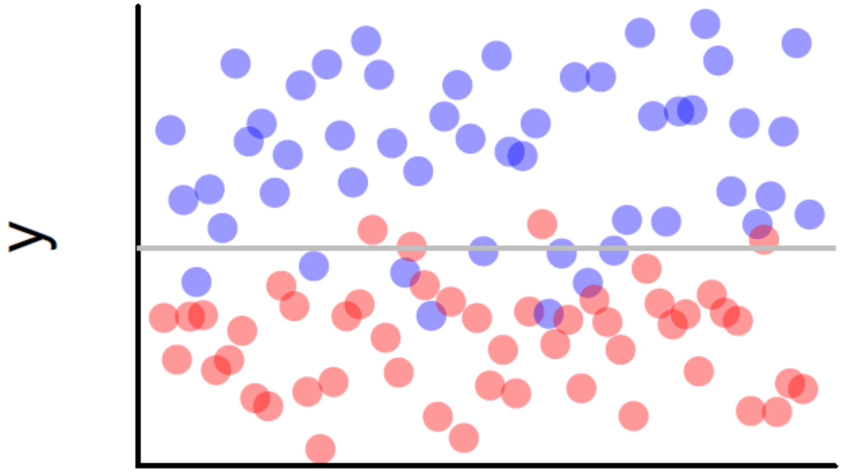
Review!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`

What does the two-level factor (t-test) look like mathematically?

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

What does the two-level factor (t-test) look like graphically?



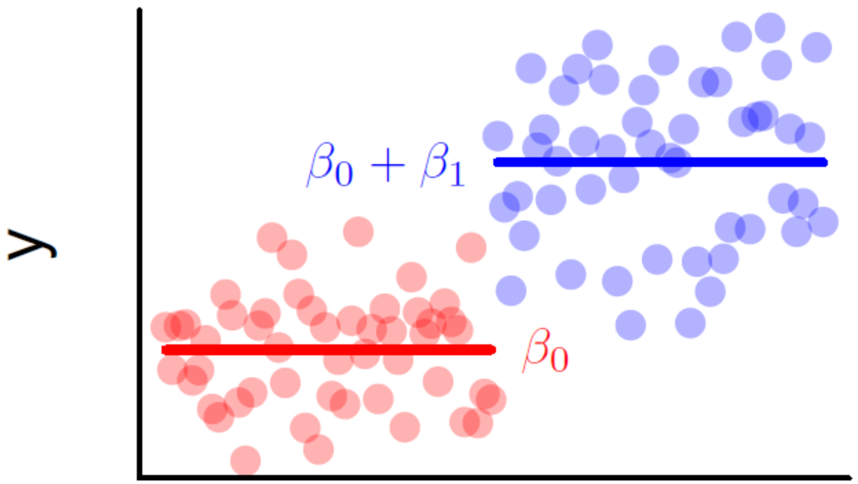
Review!

Response (Y)	Explanatory (X)	Model	In R
Continuous	None	Intercept-only/null	`lm(y~1)`
Continuous	Two-level factor	*t*-test	`lm(y~x)`

What does the two-level factor (t-test) look like mathematically?

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

What does the two-level factor (t-test) look like graphically?



For next week:

1) Finish reading Ch. 5.1 in the Zuur et al. (2007) book
2) Watch the recorded lecture and do the exercise
3) Finish the posted Week 2 lab
4) Please bring questions to class on Tuesday, as we will recap simple linear regression and move to more complex models!
5) Complete the individual assessment on Moodle by 11:55pm Monday night.