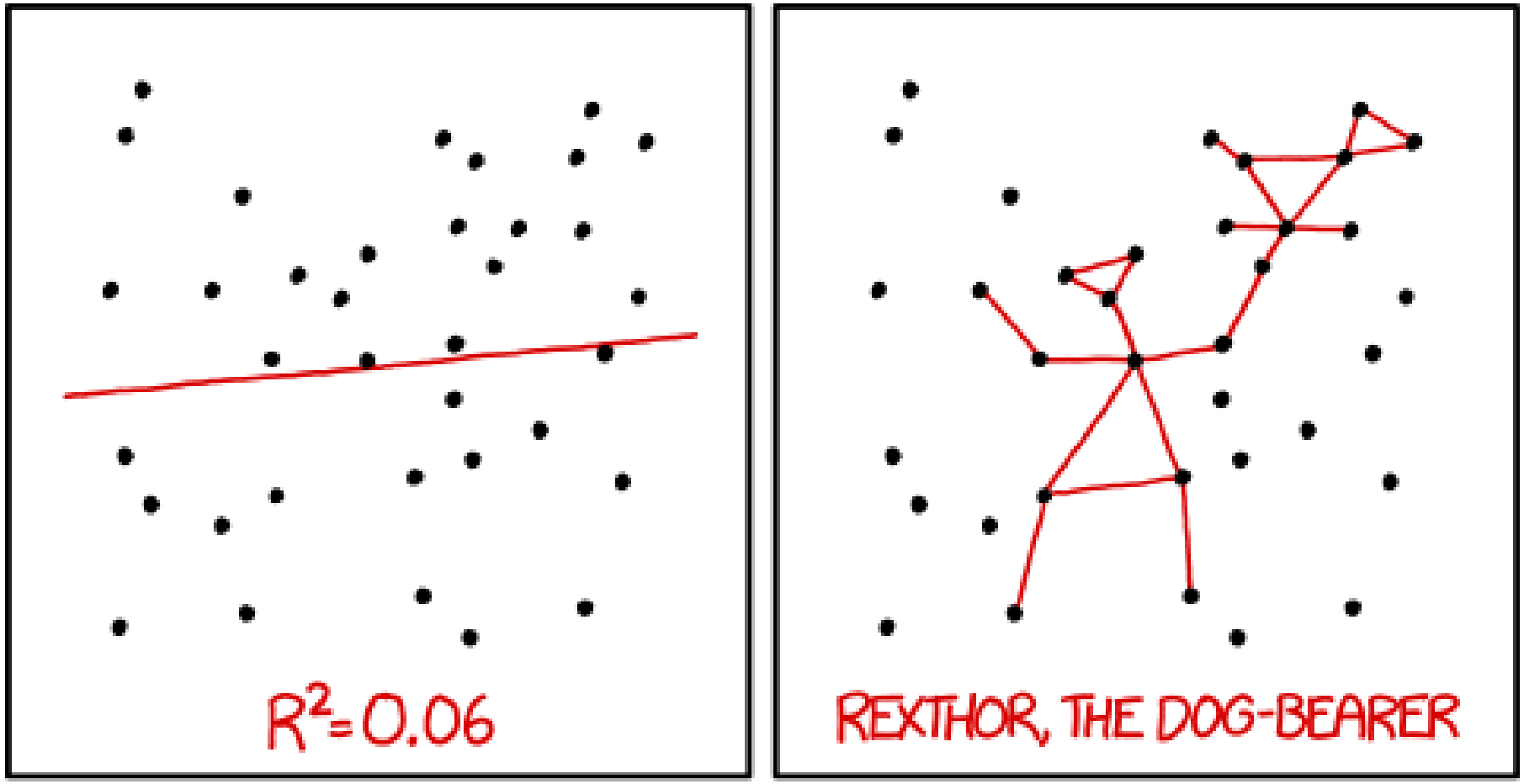
BIOL 599: Linear Models



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

https://xkcd.com/1725/

Reminders

Thursday: Paper discussion and linear regression workshop

Office hours Monday and Thursday

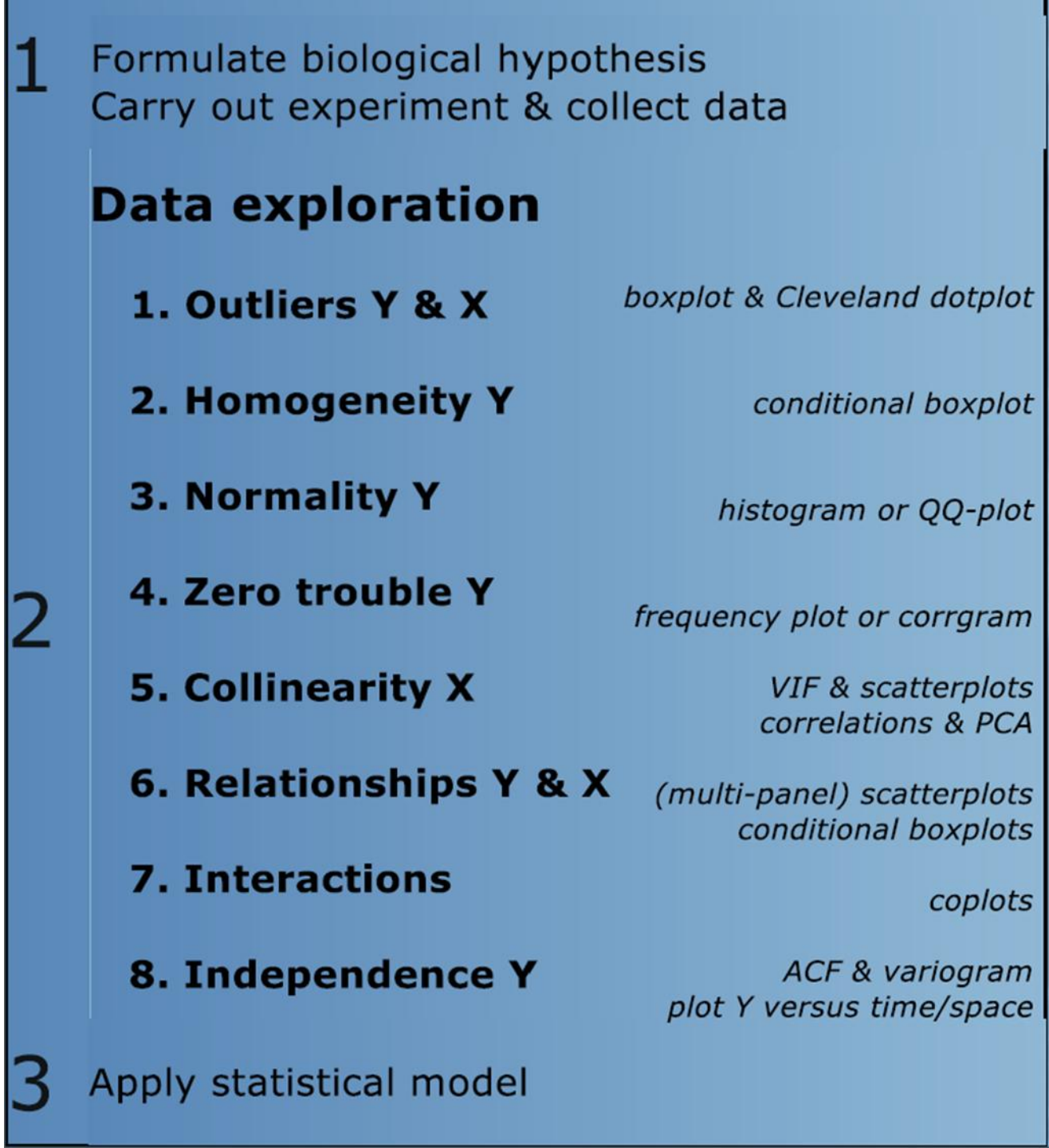
Assignment #1 DUE Friday

Review from last class

- What are the 8 steps of data exploration?

Remember:

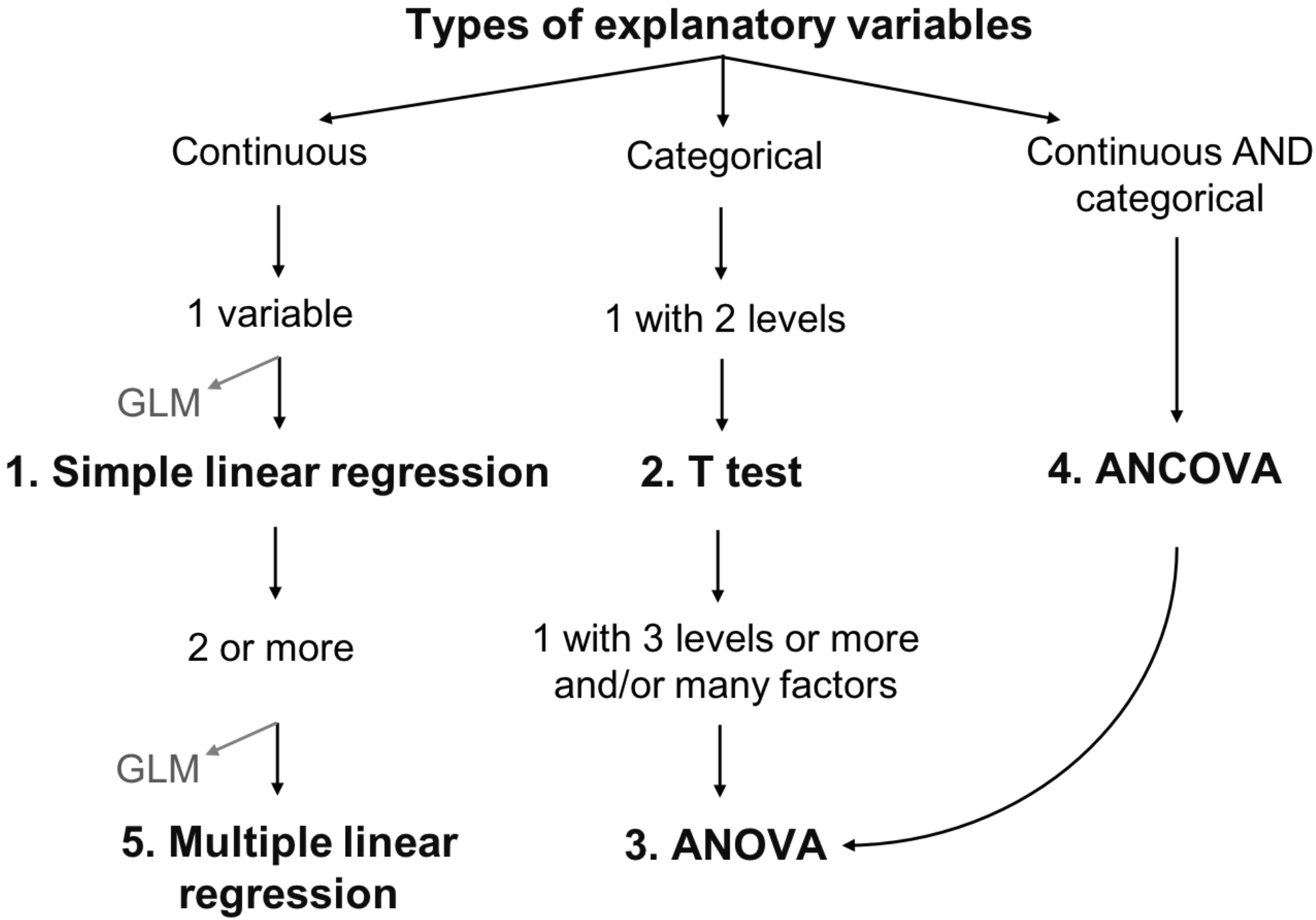
- Not every data set requires each step!
- Order also depends on data set
- Treat list as a series of questions
- ***For some analyses, assumptions can ONLY be verified after the analysis***

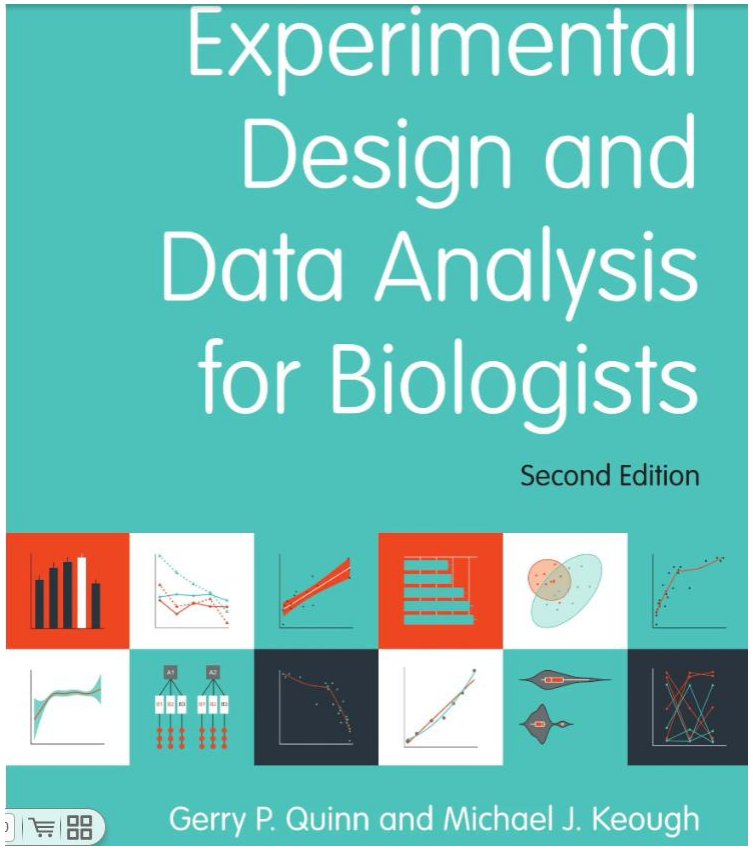


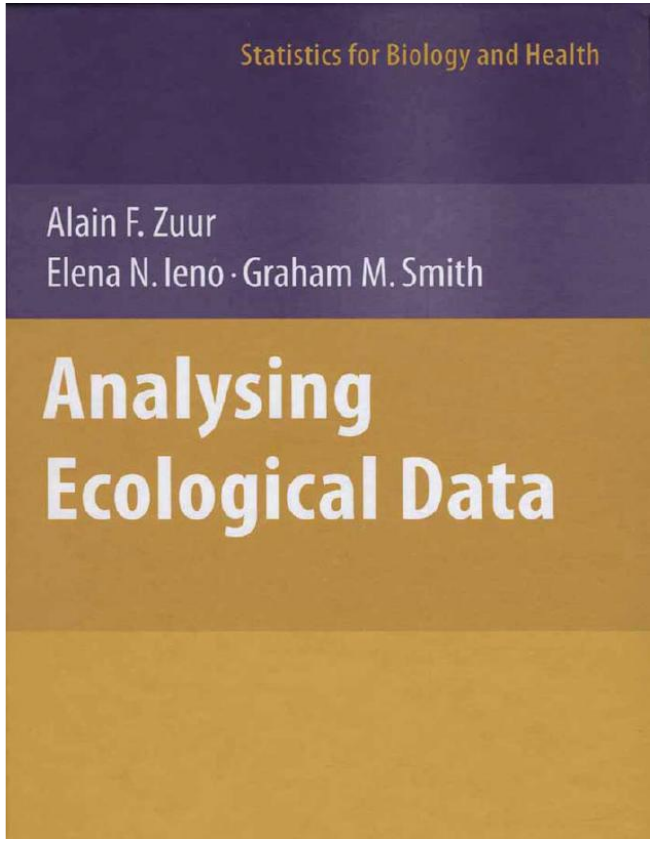
- What is a linear model?
- Simple linear regression
  - ✓ Assessing model fits and assumptions
  - ✓ Model comparison: full vs reduced
- Categorical predictor
  - ✓ T-test
  - ✓ ANOVA
- Multiple linear regression
  - ✓ Categorical and continuous predictors (ANCOVA)
  - ✓ Standardizing and centering

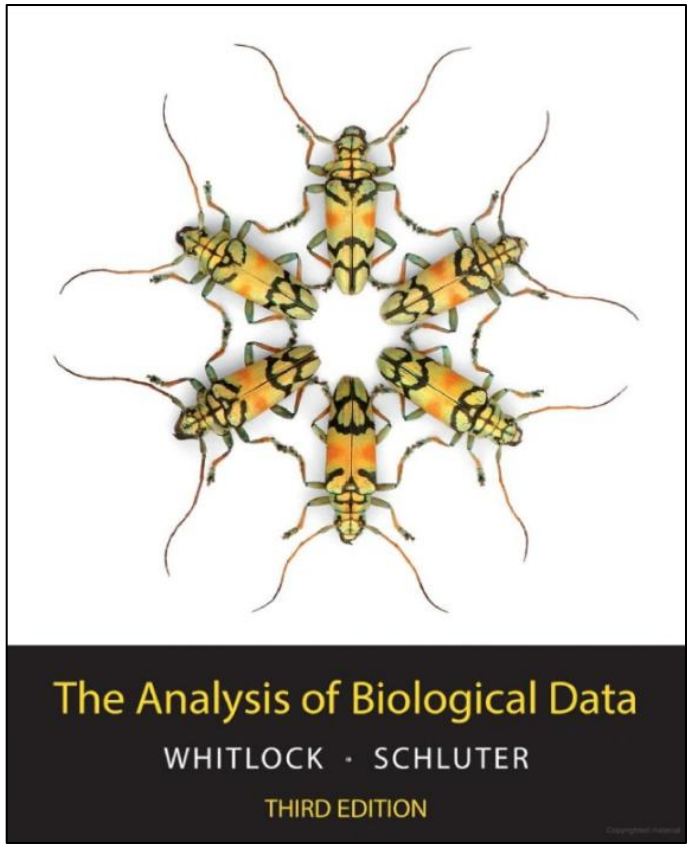
Outline

GOAL: learn the structure of a linear model and its different variants


Types of explanatory variables
Continuous
Categorical
Continuous AND categorical
1 variable
1 with 2 levels
GLM
1. Simple linear regression
2. T test
4. ANCOVA
2 or more
1 with 3 levels or more and/or many factors
GLM
5. Multiple linear regression
3. ANOVA


Experimental Design and Data Analysis for Biologists
Second Edition
Gerry P. Quinn and Michael J. Keough


Statistics for Biology and Health
Alain F. Zuur
Elena N. Ieno · Graham M. Smith
Analysing Ecological Data


The Analysis of Biological Data
WHITLOCK · SCHLUTER
THIRD EDITION

What is a linear model?

- Describes the relationship between a **response** variable and one or more **predictor** variables based on a sample

- Determines whether variables are correlated by inferring the **direction** and **strength** of a relationship, and our **confidence** in the effect size estimates.

What is a linear model?

For continuous variables:

- assesses whether the mean value of the response variable differs significantly between different values of the explanatory variables

For categorical explanatory variables:

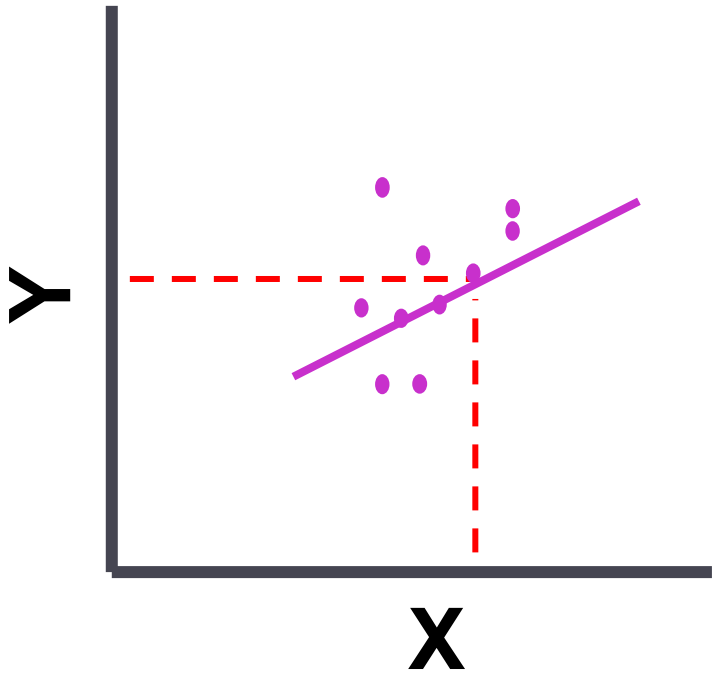
- assesses whether the mean value of the response variable differs significantly between different levels (or groups) of explanatory variables

What is a linear model?

- The line represents the assumed relationship

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

- y_i is an observation of the response y
- x_i is an observation of the predictor x
- The parameter β_0 is the **intercept**.
- The parameter β_1 quantifies the **effect** of x on y
- The residual ϵ_i represents the **unexplained** variation
- The **predicted value** of y_i is defined as: $\hat{y}_i = \beta_0 + \beta_1 \times x_i$



AIM: find the “best” estimate of the parameters β_0 and β_1

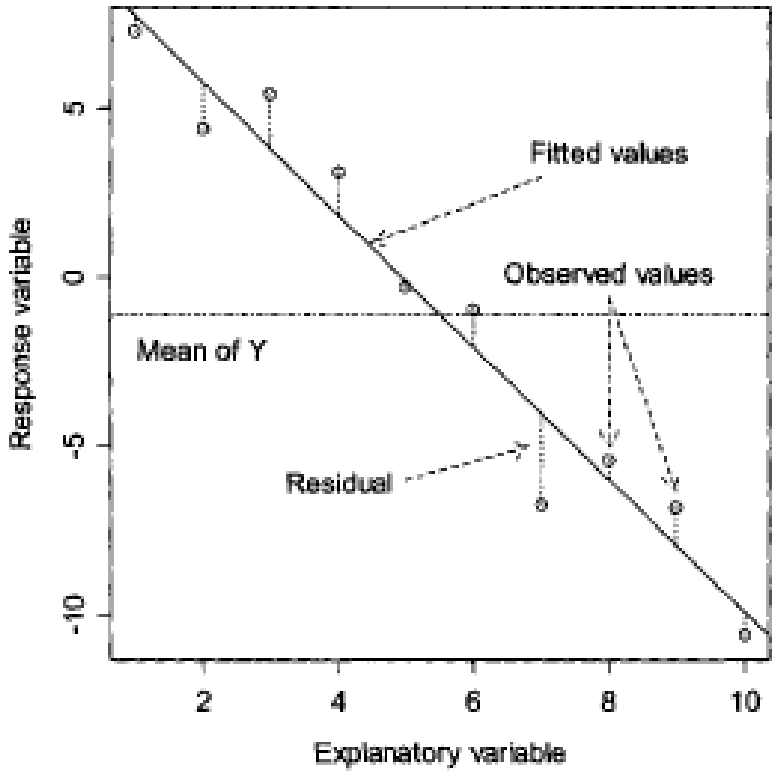
What is a linear model?

- The "best" parameters are those that minimize the sum of the squared residuals
- This method is called Ordinary Least Squares (OLS)

Table 5.1. Three variance components.

Notation	Variance in	Sum of squared deviations of	Formula
SS_{total}	Y	Observed data from the mean	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$
$SS_{regression}$	Y explained by X	Fitted values from the mean value	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
$SS_{residual}$	Y not explained by X	Observed values from fitted values	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$



What is a linear model?

ANOVA table for simple linear regression

Table 5.2. ANOVA table for simple regression model. df stands for degrees of freedom.

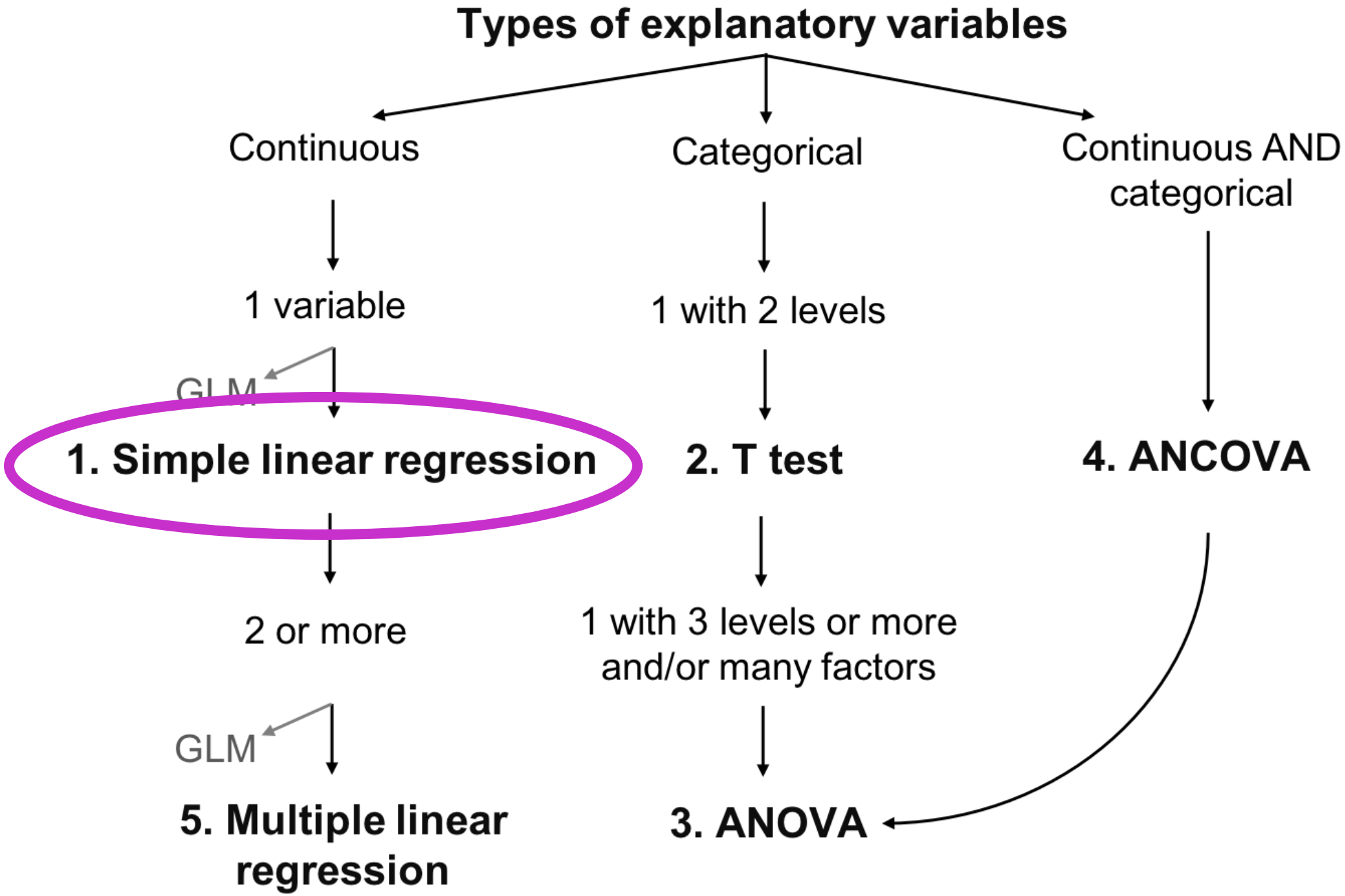
Source of variation	SS	df	MS
Regression	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	$\sum_{i=1}^{n}\frac{(\hat{Y}_i - \bar{Y})^2}{1}$
Residual	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-2$	$\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)^2}{n-2}$
Total	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$	

Null hypothesis = slope is zero

F-statistic = MSregression/MSresidual

Outline

- What is a linear model?
- Simple linear regression
  - ✓ Assessing model fits and assumptions
  - ✓ Model comparison: full vs reduced
- Categorical predictor
  - ✓ T-test
  - ✓ ANOVA
- Multiple linear regression
  - ✓ Categorical and continuous predictors (ANCOVA)
  - ✓ Standardizing and centering


Types of explanatory variables
Continuous
Categorical
Continuous AND categorical
1 variable
1 with 2 levels
GLM
1. Simple linear regression
2. T test
4. ANCOVA
2 or more
1 with 3 levels or more and/or many factors
GLM
5. Multiple linear regression
3. ANOVA

STEPS in linear regression

1.scatter plot (examine data)

plot(y ~ x, data = mydata)

2. Fit linear model

model1<- lm(y ~ x, data=mydata)

3. Look at model assumptions (diagnostics)

plot(model1), check_model(model1)

4. Extract coefficients and information from the model

summary(model1) and model1$coefficients

5. Test model fit with anova (test hypothesis)

anova(model1)

6.Model comparison between full and reduced

anova(model1, null)

7. Add model line to plot

abline() or lines() or ggplot()

visreg(), predict(), ggpredict(), or ggeffects()

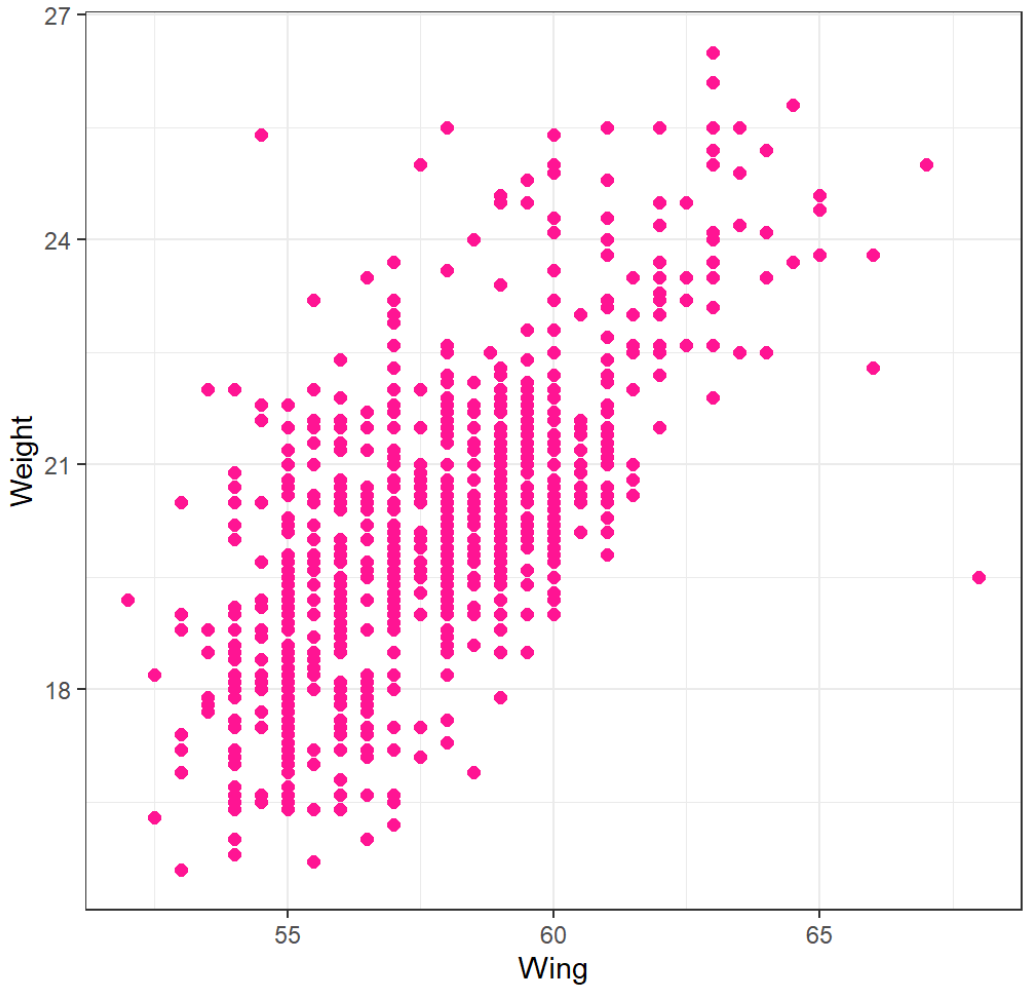
Example of linear model (both numerical X,Y)

Question: Is wing length a significant linear predictor of weight for Savannah sparrows?

Hypothesis: For Savannah sparrows, the **wing length (mm)** of an individual influences the **weight (grams)** of an individual

```
> spar=read.csv(file="sparrows.csv")
> head(spar)
  Species    Sex Wing Tarsus Head Culmen Nalospi Weight Observer Age
1    SSTS   Male 58.0   21.7 32.7   13.9    10.2   20.3        2   0
2    SSTS Female 56.5   21.1 31.4   12.2    10.1   17.4        2   0
3    SSTS   Male 59.0   21.0 33.3   13.8    10.0   21.0        2   0
4    SSTS   Male 59.0   21.3 32.5   13.2     9.9   21.0        2   0
5    SSTS   Male 57.0   21.0 32.5   13.8     9.9   19.8        2   0
6    SSTS Female 57.0   20.7 32.5   13.3     9.9   17.5        2   0
```

1. Plot and Examine Data




```
ggplot(spar, aes(y=Weight, x=Wing)) + geom_point(size=2, color="deeppink") + theme_bw()
```

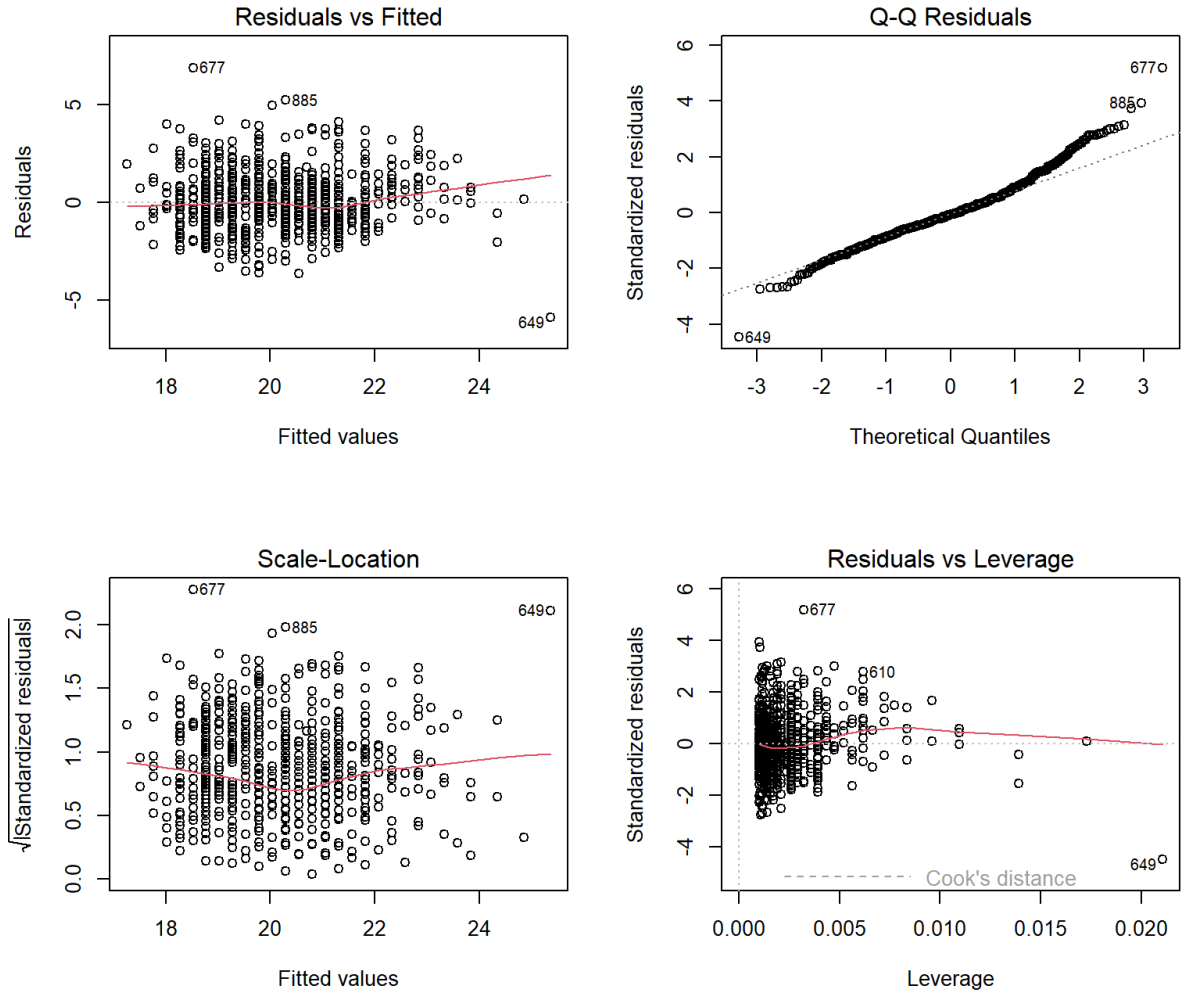
2. Fit a linear model with lm()

We are modeling **only fixed** factors with lm()

```
model1<- lm(y ~ x, data=mydata)
```

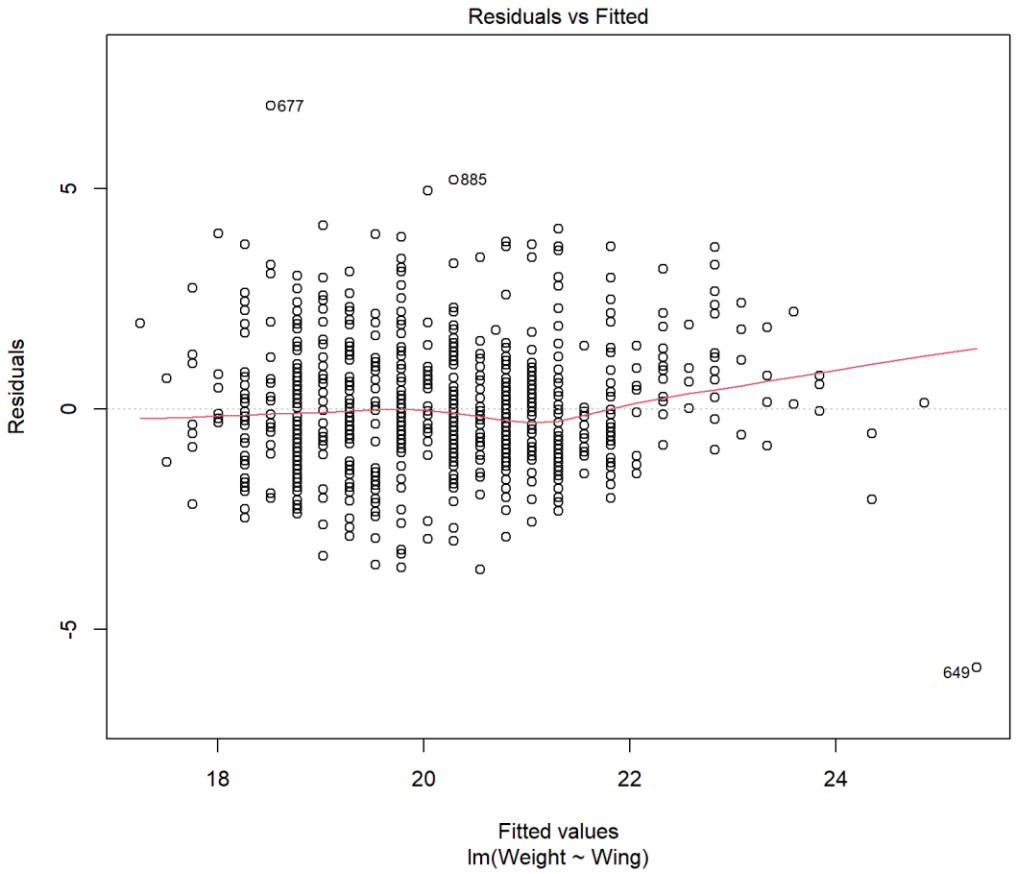
```
model1=lm(data=spar, Weight~Wing)
```

3. Verify assumptions



How do we interpret these plots?

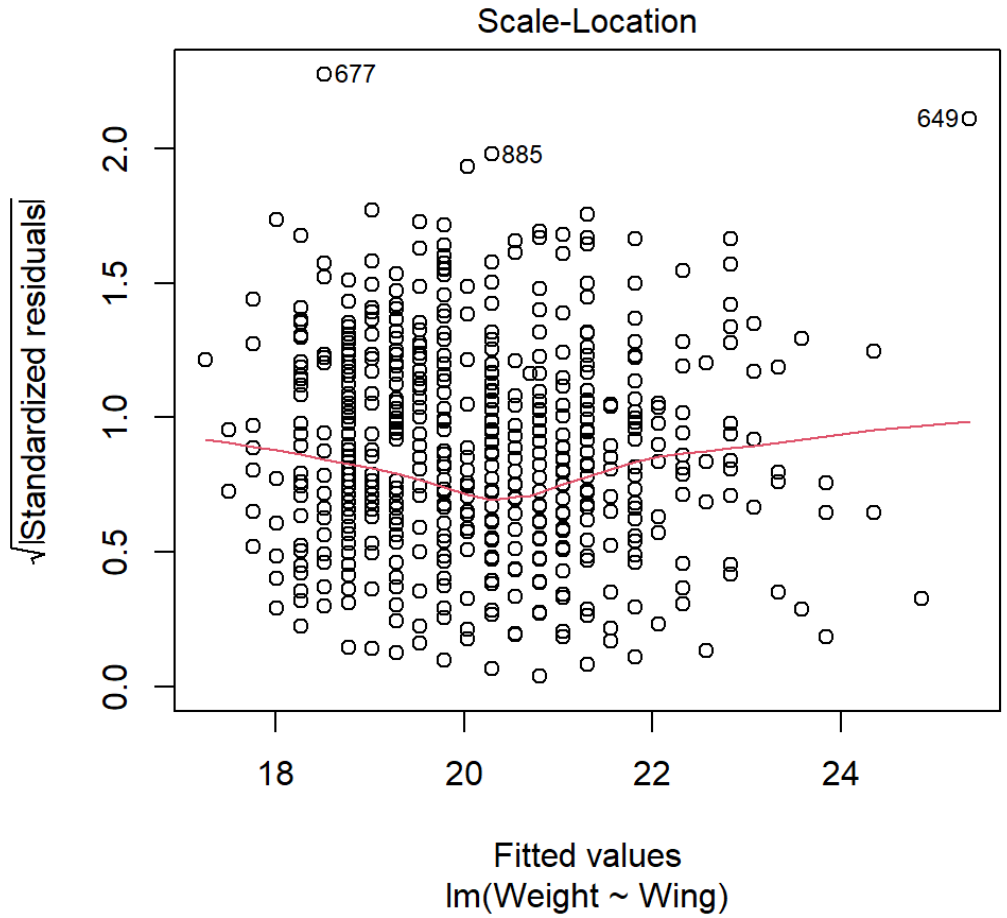
Residuals vs Fitted → *pattern check* (linearity)



What we hope to see: Random scatter, no pattern

Why: Shows residuals are *independent* and *identically distributed*

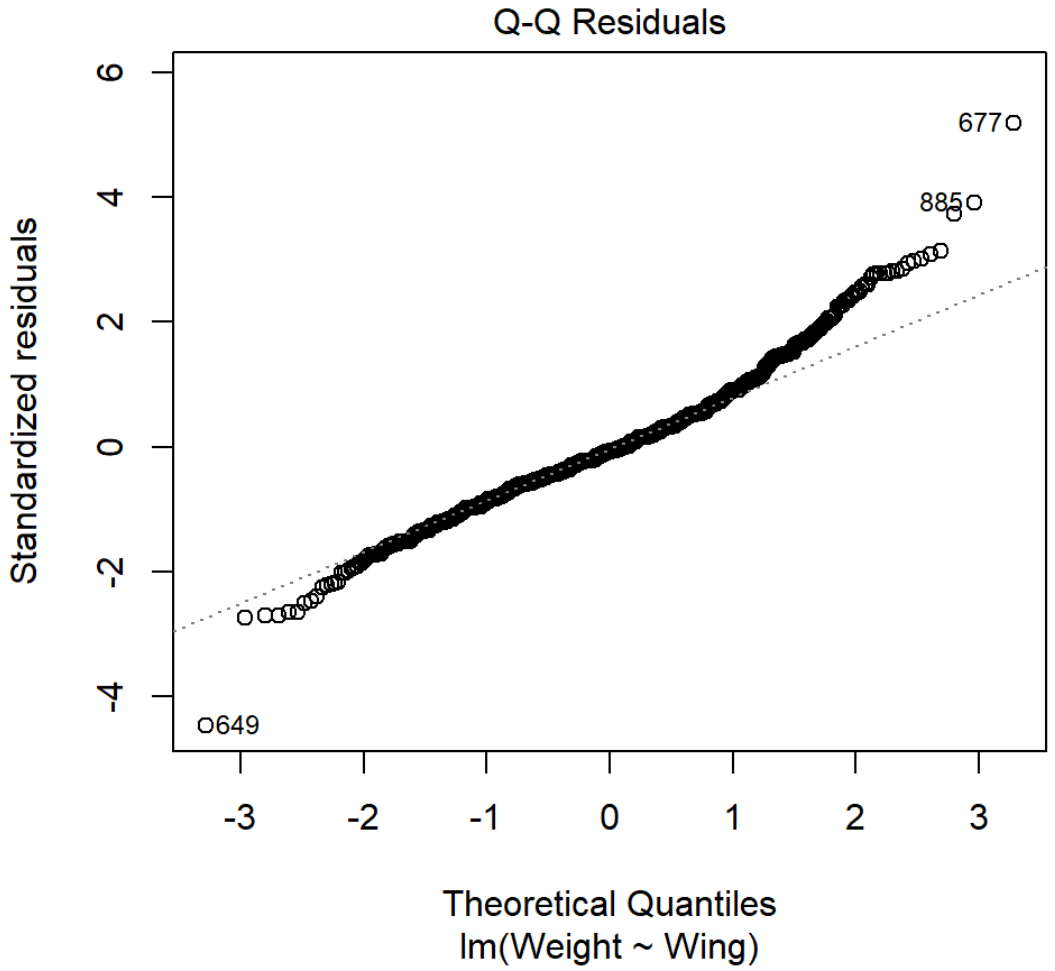
Scale–Location → *spread check* (equal variance).



What we hope to see: Random scatter, no pattern, no “outliers”

Why: Violations of assumptions are sometimes easier to detect than in the first plot, especially when the predictor is not uniformly distributed.

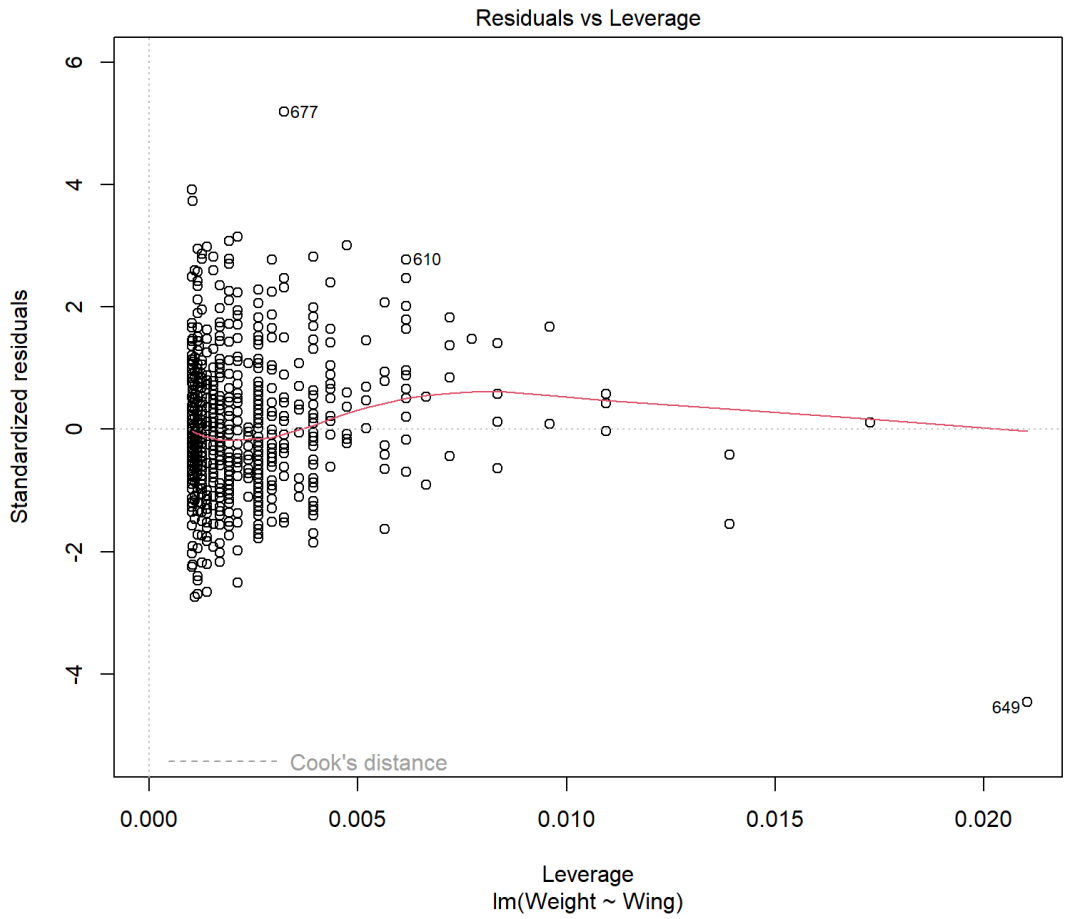
Q–Q Plot → “Shape check” (normality).



What we hope to see: Points clearly on the 1:1 line

Why: Compares the distribution (quantiles) of the residuals with a standard normal distribution

Residuals vs Leverage → Influence (outliers)

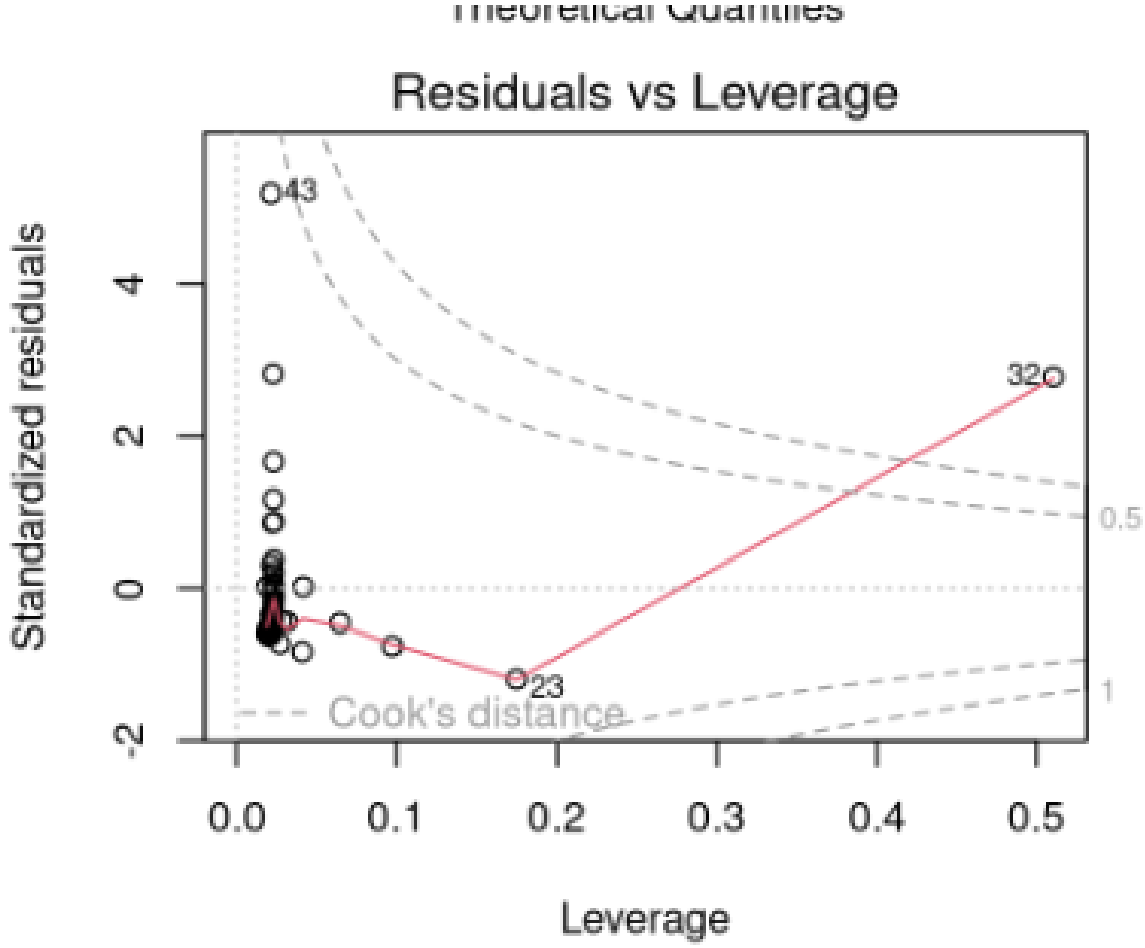


What we hope to see: No leverage points with high influence

Why: Leverage points may or may not have a high influence on the regression

“Cook’s distance”

- Combines leverage + residual size
- Tells us how much the regression line would change if we removed a point
- Rule of thumb: values > 1 are worrisome
- High leverage alone ≠ always influential
- Large residual alone ≠ always influential.



A few more notes on checking assumptions

- As well as normality, homogeneity and independence, you should also check for residual patterns in the data. This assesses model misspecification and model fit.
- Plot residuals against each explanatory variable to check that no clear patterns are shown by the plotted residuals.
  - Continuous predictors: the residuals should be scattered equally across the whole graph.
  - Categorical predictors: equal variance of residuals between groups

What if my data violates assumptions?

If the plotted residuals show an obvious non-random structure, several options are available:

1. Apply a transformation.
2. Add other explanatory variables.
3. Add interactions.
4. Add non-linear terms of the explanatory variables (e.g. quadratic terms).
5. Use smoothing techniques like additive modeling (GAM)
6. Allow for different spread using generalized least squares (GLS)
7. Apply mixed modeling (LMM)

"In none of these [35] real data sets could we find a non-trivial example for a linear regression model for which all assumptions held." Zuur 2007

STEPS in linear regression

1.scatter plot (examine data)

plot(y ~ x, data = mydata)

2. Fit linear model

model1<- lm(y ~ x, data=mydata)

3. Look at model assumptions (diagnostics)

plot(model1), check_model(model1)

4. Extract coefficients and information from the model

summary(model1) and model1$coefficients

5. Test model fit with anova (test hypothesis)

anova(model1)

6.Model comparison between full and reduced

anova(model1, null)

7. Add model line to plot

abline() or lines() or ggplot()

visreg(), predict(), ggpredict(), or ggeffects()

4. Use summary() to get parameter estimates

```
summary(model1)
library(sjPlot)
tab_model(model1)
```

```
Call:
lm(formula = Weight ~ Wing, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8673 -0.7872 -0.1017  0.6910  6.8811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.12820    1.07459  -8.495   <2e-16 ***
Wing         0.50729    0.01856  27.339   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.329 on 977 degrees of freedom
Multiple R-squared:  0.4334,    Adjusted R-squared:  0.4328
F-statistic: 747.4 on 1 and 977 DF,  p-value: < 2.2e-16
```

	Weight		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-9.13	-11.24 – -7.02	**<0.001**
Wing	0.51	0.47 – 0.54	**<0.001**
Observations	979		
R^2 / R^2 adjusted	0.433 / 0.433		

Useful for more complex models

But ignore the tests for significance
(we will use anova later)

Discussion:
what do the parameter estimates tell us?

```
Call:
lm(formula = Weight ~ Wing, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8673 -0.7872 -0.1017  0.6910  6.8811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.12820    1.07459  -8.495   <2e-16 ***
Wing         0.50729    0.01856  27.339   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.329 on 977 degrees of freedom
Multiple R-squared:  0.4334,    Adjusted R-squared:  0.4328
F-statistic: 747.4 on 1 and 977 DF,  p-value: < 2.2e-16
```

	Weight		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-9.13	-11.24 – -7.02	**<0.001**
Wing	0.51	0.47 – 0.54	**<0.001**
Observations	979		
R^2 / R^2 adjusted	0.433 / 0.433		

- Coefficients table: test slope = 0 using *t-statistic*
- Overall model test: shown by *F-statistic*
- In simple linear regression: $t^2=F$

5. Test the hypothesis with anova(model1)

- Null hypothesis is that slope=0 (that there is no line)
- **anova(model1) asks, "Is this model linear?"**
- Yields an anova table

```
Analysis of Variance Table

Response: Weight
           Df Sum Sq Mean Sq F value    Pr(>F)
Wing        1 1320.2 1320.17  747.39 < 2.2e-16 ***
Residuals 977 1725.7    1.77
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test of null hypothesis that slope $\beta_1 = 0$

6. Model comparison between a full and reduced model

- anova() on 2 models compares the model fits with an F-test

****must be comparing a reduced vs. full model otherwise test is invalid****

- The full model contains the term of interest and the reduced model leaves it out.
- The reduced and full model **only differ by the term of interest**
- Sometimes termed hierarchically nested

This concept will be very important when we compare more complex models

6. Model comparison between a full and reduced model

```r
#fit a  reduced model (with intercept only, no slope)
null=lm(data=spar, Weight~1)

#fit a full model with intercept and wing
model1=lm(data=spar, Weight~Wing)

#compare the reduce (null) vs full model
#This does an F-tst --ANOVA table
anova(null, model1)
```

The anova function finds the factors that have been removed in the reduced model as compared to the full model, and tests whether those removed factors are different from zero.

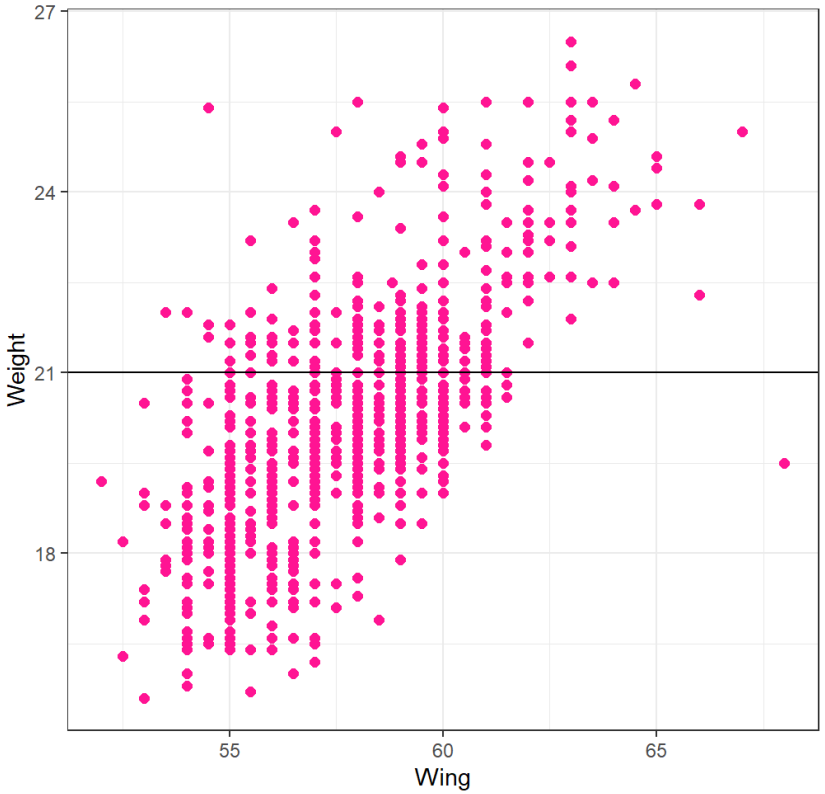
- Null: removed factors are not different from 0 - full model is not better
- Alt: removed factors are different from 0, full model is better
- $p < 0.05$ => choose full model
- $p > 0.05$ => choose reduced model

Visually, how R compares models

anova(null,model1)

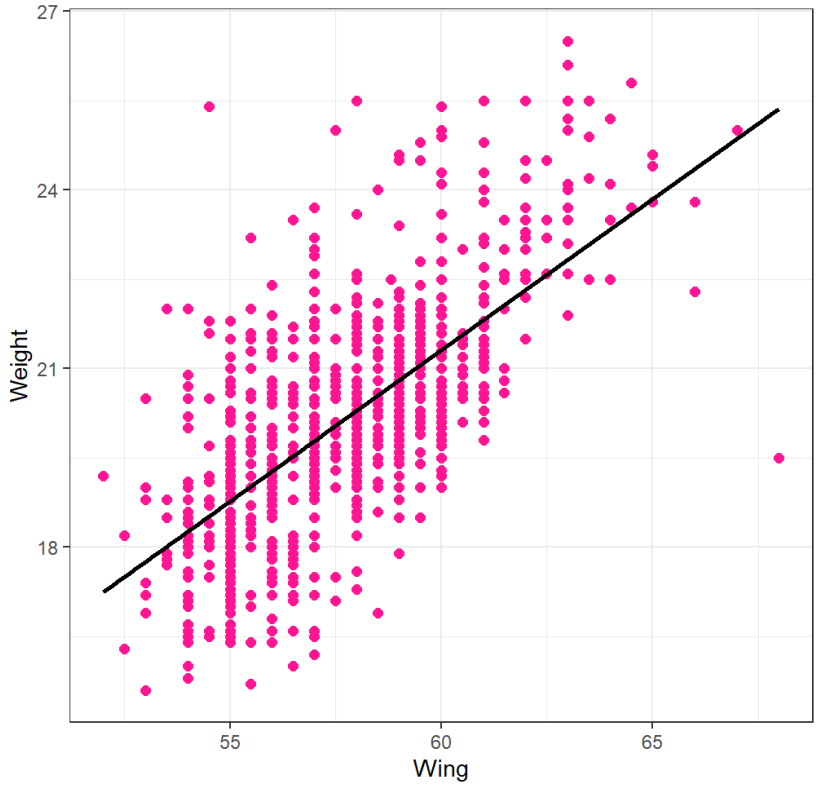
Weight ~ 1

reduced model (fits only an intercept)



Weight ~ Wing

full model (intercept and slope)



6. Model comparison between a full and reduced model

anova(null,model1) produces an F-test R output

```
Analysis of Variance Table

Model 1: Weight ~ 1
Model 2: Weight ~ Wing
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    978 3045.9
2    977 1725.7  1    1320.2 747.39 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value<0.05 means that removed factors are different from 0, full model is better

Don’t mix up the anovas

anova (model1)→*Type I*

tests hypothesis: “Is it linear?”

NOTE: this is sequential with multiple variables – order matters

anova(null,model1)→ compares full vs reduced models

“Is full model better than the reduced model”

- *Type II* – use with multiple predictors, no interactions
- *Type III* – use with interactions, but interpret main effects cautiously

summary(model1)

```
Call:
lm(formula = Weight ~ Wing, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8673 -0.7872 -0.1017  0.6910  6.8811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.12820    1.07459  -8.495   <2e-16 ***
Wing         0.50729    0.01856  27.339   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.329 on 977 degrees of freedom
Multiple R-squared:  0.4334,    Adjusted R-squared:  0.4328
F-statistic: 747.4 on 1 and 977 DF,  p-value: < 2.2e-16
```

anova(model1)

```
Analysis of Variance Table

Response: Weight
           Df Sum Sq Mean Sq F value    Pr(>F)
Wing        1 1320.2 1320.17  747.39 < 2.2e-16 ***
Residuals 977 1725.7    1.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

anova(null, model1)

```
Analysis of Variance Table

Model 1: Weight ~ 1
Model 2: Weight ~ Wing
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    978 3045.9
2    977 1725.7  1    1320.2 747.39 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

STEPS in linear regression

1.scatter plot (examine data)

plot(y ~ x, data = mydata)

2. Fit linear model

model1<- lm(y ~ x, data=mydata)

3. Look at model assumptions (diagnostics)

plot(model1), check_model(model1)

4. Extract coefficients and information from the model

summary(model1) and model1$coefficients

5. Test model fit with anova (test hypothesis)

anova(model1)

6.Model comparison between full and reduced

anova(model1, null)

7. Add model line to plot

abline() or lines() or ggplot()

visreg(), predict(), ggpredict(), or ggeffects()

7. Plot the fitted model line over the scatterplot

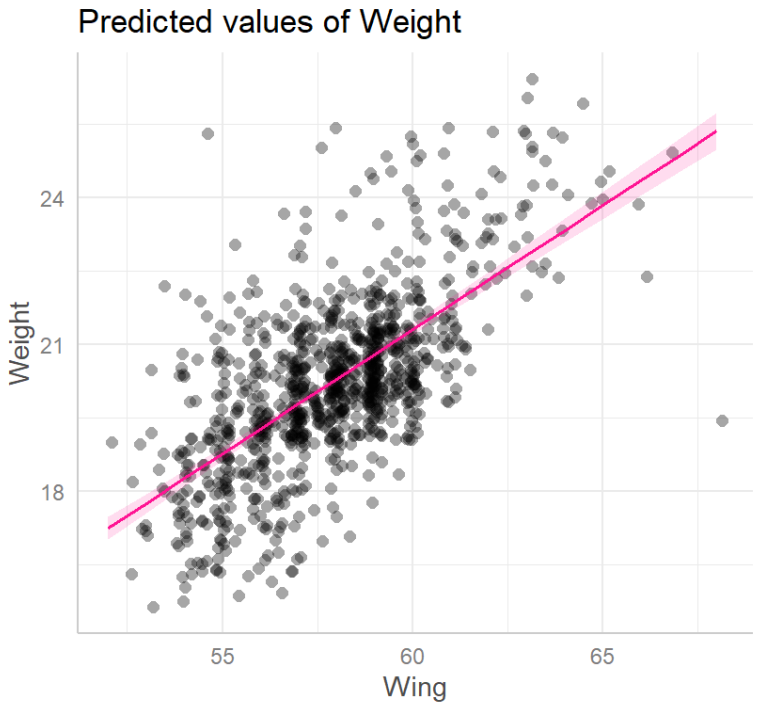
- lines()
- abline(model1)
- geom_abline() or geom_smooth()
- CI and model line with visreg() or ggpredict()

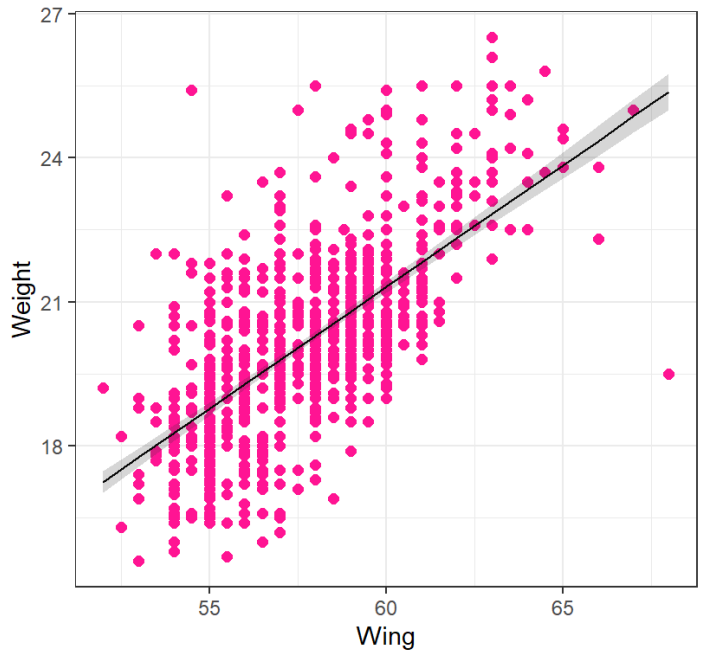
You have lots of options in how to plot model lines

Example with ggpredict()

- Scatterplot, model line, and 95%CI – useful for more complicated models

```
#default
m=ggpredict(model = model1, terms = c("Wing"))
plot(m, add.data=TRUE, color="deeppink")

#make it pretty :)
g=ggplot() + geom_point(data=spar, aes(y=Weight, x=Wing), color="deeppink", size=2) + theme_bw()
g=g+ geom_line(data=m, aes(y=predicted, x=x), color="black")
g=g+geom_ribbon(data=m, aes(x=x, ymin=conf.low, ymax=conf.high), alpha=0.2)
g
```



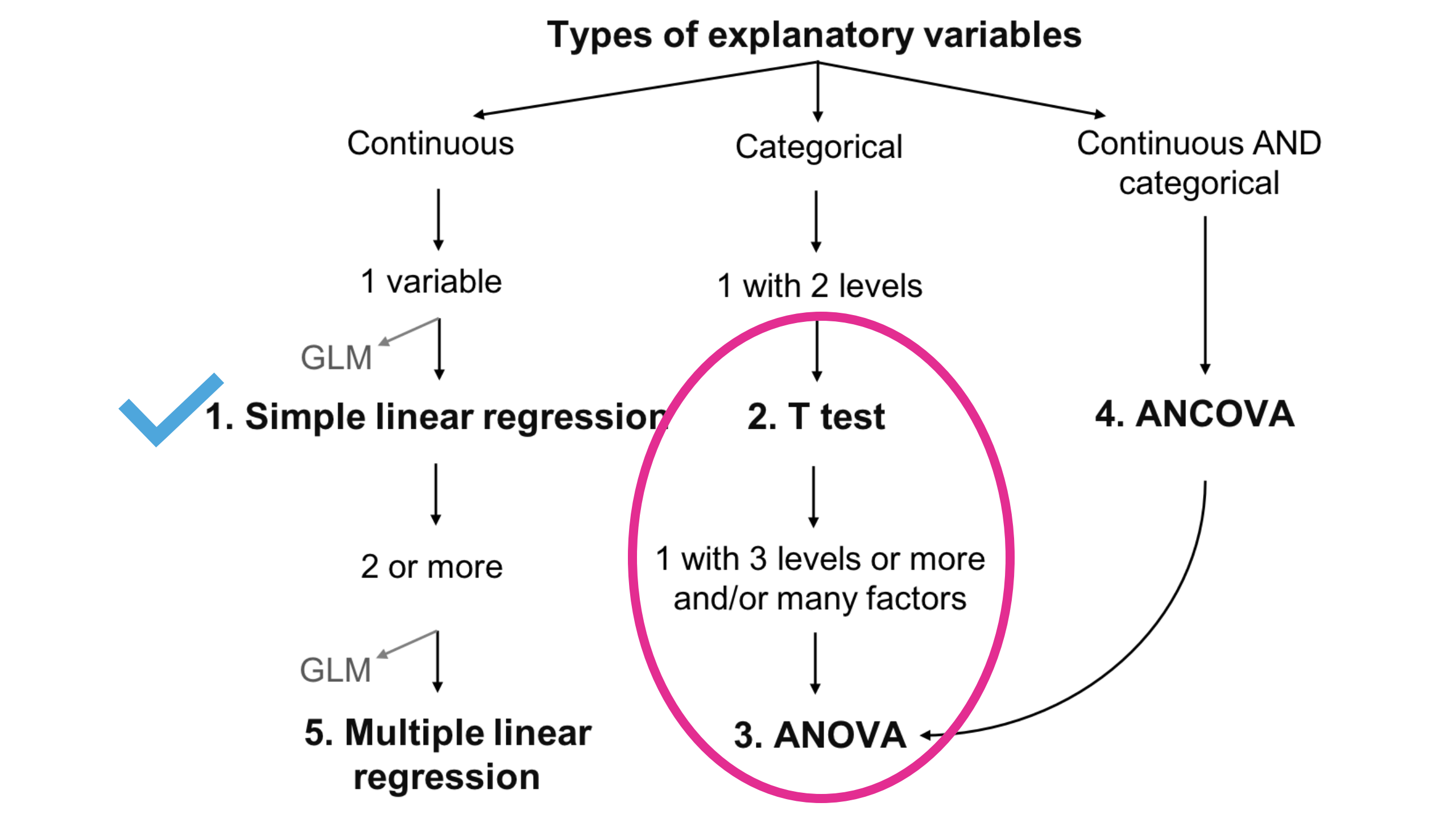


Linear Regression: muddy vs. clear?

- Share one clear point you feel confident about from linear regression.
  - Share one muddy point that still feels confusing or uncertain.

Outline

- What is a linear model?
- Simple linear regression
  - ✓ Assessing model fits and assumptions
  - ✓ Model comparison: full vs reduced
- Categorical predictor
  - ✓ T-test
  - ✓ ANOVA
- Multiple linear regression
  - ✓ Categorical and continuous predictors (ANCOVA)
  - ✓ Standardizing and centering


Types of explanatory variables
Continuous
Categorical
Continuous AND categorical
1 variable
1 with 2 levels
GLM
1. Simple linear regression
2. T test
4. ANCOVA
2 or more
1 with 3 levels or more and/or many factors
GLM
5. Multiple linear regression
3. ANOVA

Analysis of Variance (ANOVA) and t-test

Y: Response variable is **continuous**.
X: Explanatory variable(s) is **categorical** and have **two** (*t*-test) or **more** levels (ANOVA).

***t*-test** tests whether the means of the groups (or populations) are equal

H0: $\mu 1=\mu 2$

H1: $\mu 1\neq\mu 2$

ANOVA tests whether two or more means of the groups (or populations) are equal.

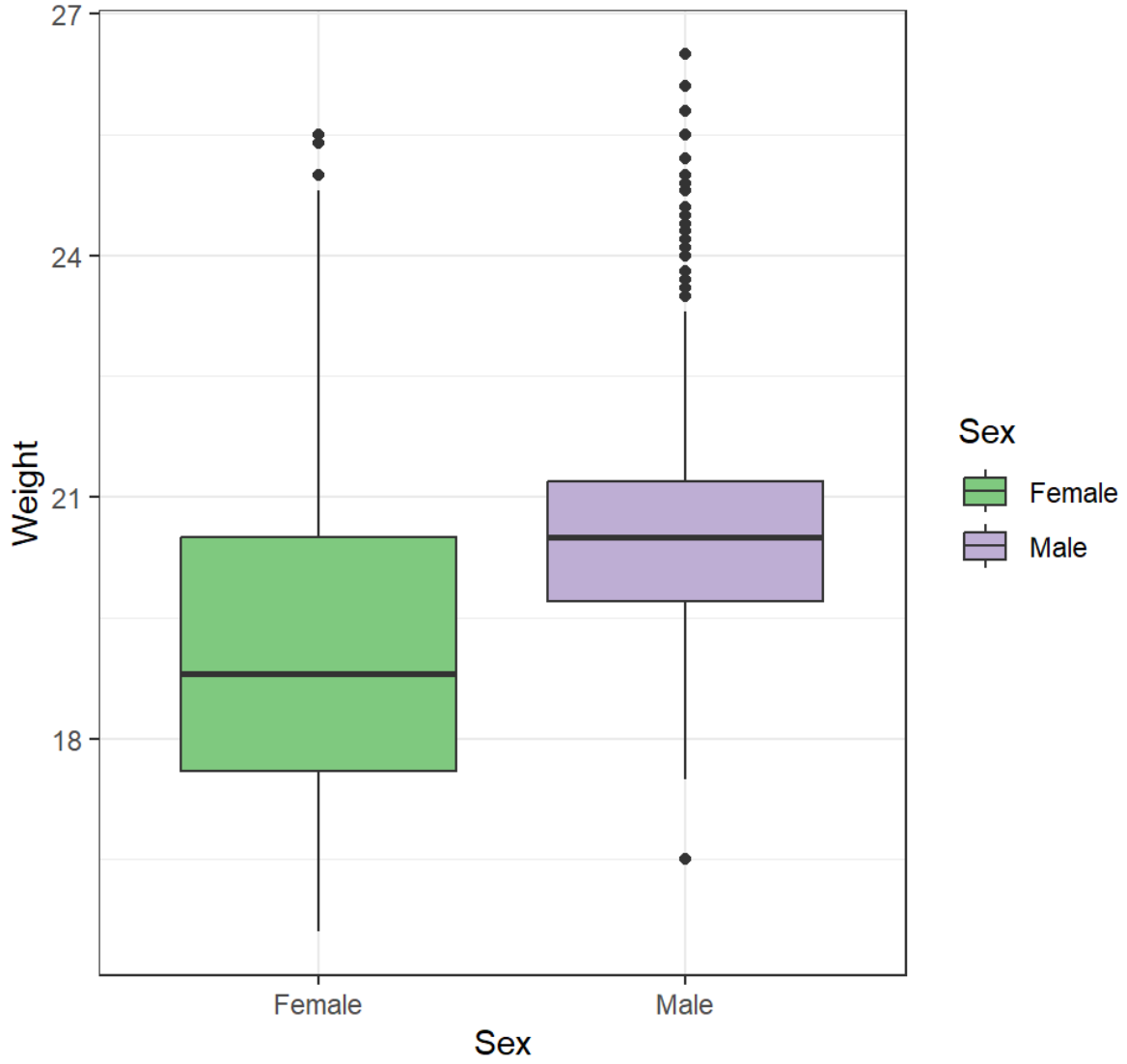
H0: $\mu 1=\mu 2=\ldots=\mu n$

H1: At least one of the group means is different across the groups

T-test

Question: Does the weight of Savannah sparrows differ between male and female sparrows?

Hypothesis: For Savannah sparrows, the **weight (grams)** will differ between male and female individuals



T-test

```
> model2<-lm(data=spar, Weight~Sex)
> summary(model2)

Call:
lm(formula = Weight ~ Sex, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1486 -1.0486 -0.2187  0.7514  6.2813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2187     0.0965  199.15   <2e-16 ***
SexMale       1.4299     0.1149   12.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.641 on 977 degrees of freedom
Multiple R-squared:  0.1367,    Adjusted R-squared:  0.1358
F-statistic: 154.7 on 1 and 977 DF,  p-value: < 2.2e-16
```

```
> t.test(data=spar, Weight~Sex)

        Welch Two Sample t-test

data:  Weight by Sex
t = -10.79, df = 410.5, p-value < 2.2e-16
alternative hypothesis: true difference in means between
equal to 0
95 percent confidence interval:
 -1.690353 -1.169378
sample estimates:
mean in group Female   mean in group Male
            19.21869             20.64855
```

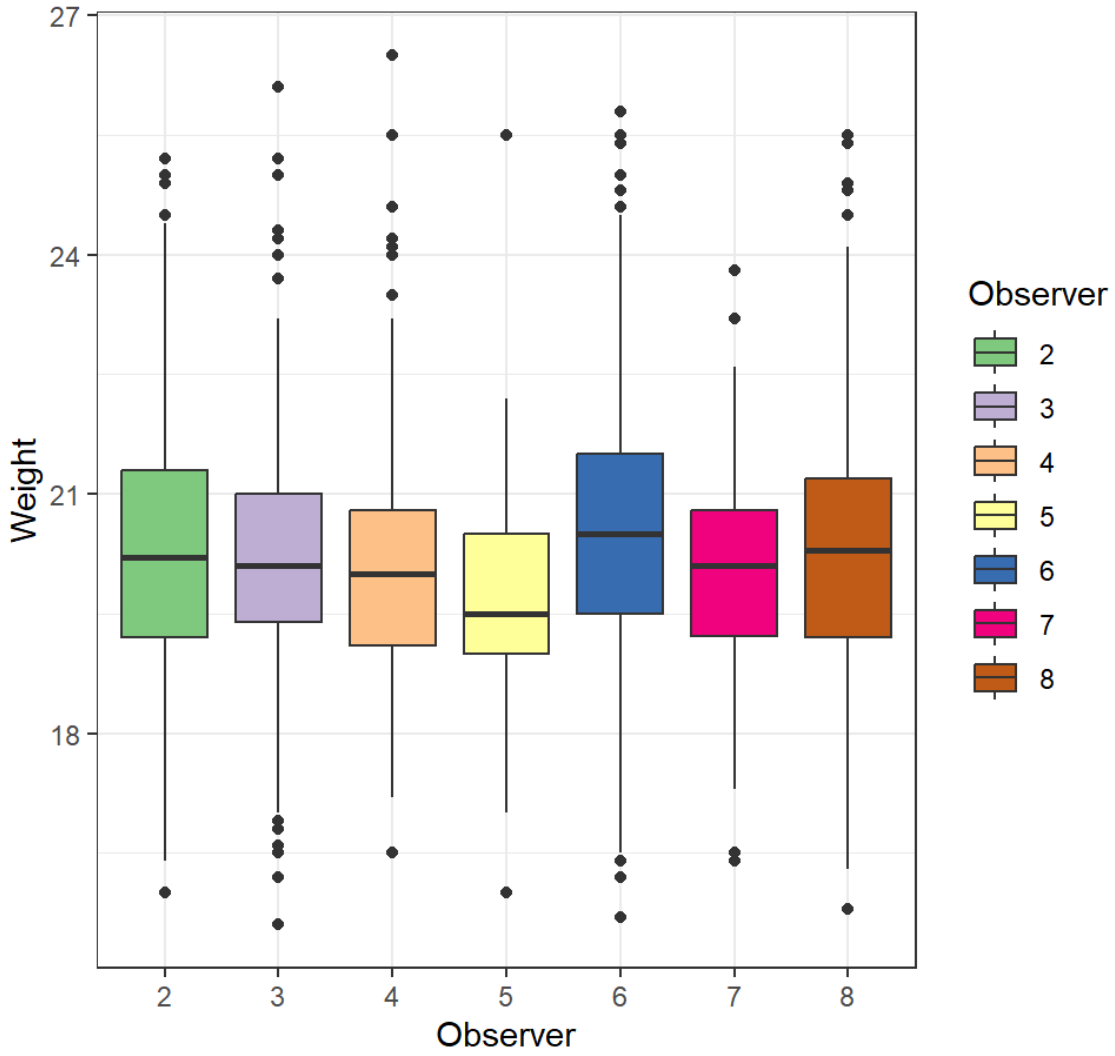
```
> anova(model2)
Analysis of Variance Table

Response: Weight
           Df  Sum Sq Mean Sq F value    Pr(>F)
Sex         1  416.44  416.44  154.73 < 2.2e-16 ***
Residuals 977 2629.46    2.69
```

```
> null<-lm(data=spar, Weight~1)
> anova(model2, null)
Analysis of Variance Table

Model 1: Weight ~ Sex
Model 2: Weight ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    977 2629.5
2    978 3045.9 -1   -416.44 154.73 < 2.2e-16 ***
```

ANOVA

Question: Does the weight of Savannah sparrows vary across different observers?

Hypothesis: For Savannah sparrows, the **weight (grams)** of an individual will vary across observers.



ANOVA

```
> model4<-lm(data=spar, Weight~Observer)
> summary(model4)

Call:
lm(formula = Weight ~ Observer, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7377 -0.9793 -0.0377  0.9207  6.2602

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.237710   0.102350 197.730   <2e-16 ***
Observer3   -0.058457   0.152923  -0.382   0.7023
Observer4    0.002049   0.218999   0.009   0.9925
Observer5   -0.531044   0.282160  -1.882   0.0601 .
Observer6    0.199952   0.175153   1.142   0.2539
Observer7   -0.195710   0.269630  -0.726   0.4681
Observer8    0.053115   0.197533   0.269   0.7881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.764 on 972 degrees of freedom
Multiple R-squared:  0.007147,  Adjusted R-squared:  0.001019
F-statistic: 1.166 on 6 and 972 DF,  p-value: 0.3221
```

```
> summary(aov1)
             Df Sum Sq Mean Sq F value Pr(>F)
Observer      6   21.8   3.628   1.166  0.322
Residuals   972 3024.1   3.111
```

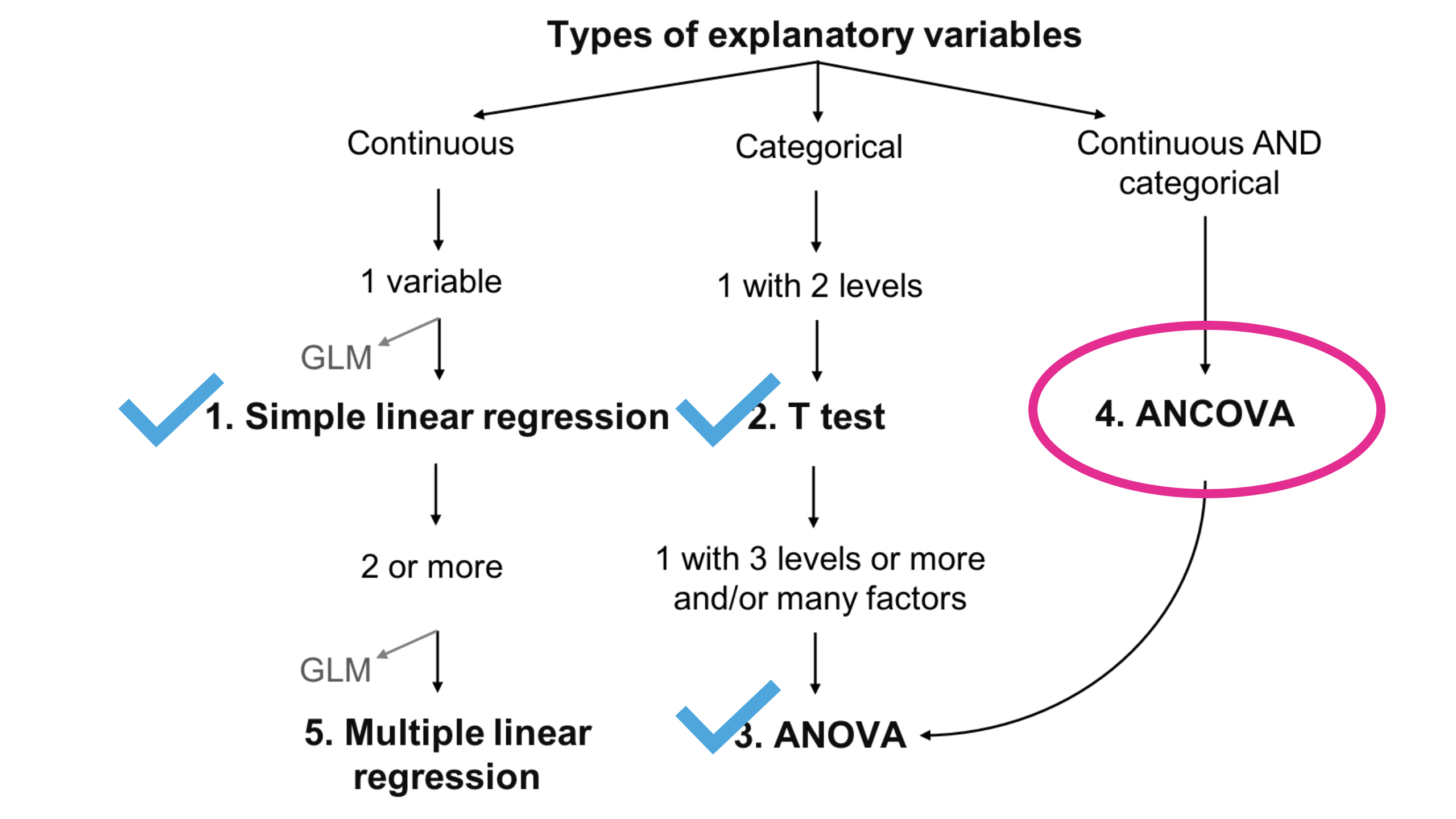
ANOVA – posthoc comparisons

- emmeans package
- Default is “tukey” adjustment to control for family-wise error rate (FWER)
- Adjusts p-values so that the overall chance of a false positive across all pairwise comparisons stays at 5%

```
> library(emmeans)
> emm.fit<- emmeans(model4, "Observer", data=spar)
> pairs(emm.fit)
 contrast              estimate    SE  df t.ratio p.value
 Observer2 - Observer3  0.05846 0.153 972   0.382  0.9998
 Observer2 - Observer4 -0.00205 0.219 972  -0.009  1.0000
 Observer2 - Observer5  0.53104 0.282 972   1.882  0.4928
 Observer2 - Observer6 -0.19995 0.175 972  -1.142  0.9151
 Observer2 - Observer7  0.19571 0.270 972   0.726  0.9910
 Observer2 - Observer8 -0.05312 0.198 972  -0.269  1.0000
 Observer3 - Observer4 -0.06051 0.224 972  -0.270  1.0000
 Observer3 - Observer5  0.47259 0.286 972   1.650  0.6498
 Observer3 - Observer6 -0.25841 0.182 972  -1.420  0.7910
 Observer3 - Observer7  0.13725 0.274 972   0.501  0.9988
 Observer3 - Observer8 -0.11157 0.204 972  -0.548  0.9981
 Observer4 - Observer5  0.53309 0.327 972   1.633  0.6612
 Observer4 - Observer6 -0.19790 0.240 972  -0.824  0.9825
 Observer4 - Observer7  0.19776 0.316 972   0.626  0.9960
 Observer4 - Observer8 -0.05107 0.257 972  -0.199  1.0000
 Observer5 - Observer6 -0.73100 0.299 972  -2.446  0.1809
 Observer5 - Observer7 -0.33533 0.362 972  -0.925  0.9685
 Observer5 - Observer8 -0.58416 0.313 972  -1.869  0.5015
 Observer6 - Observer7  0.39566 0.287 972   1.378  0.8136
 Observer6 - Observer8  0.14684 0.221 972   0.665  0.9944
 Observer7 - Observer8 -0.24883 0.301 972  -0.826  0.9823
```

Outline

- What is a linear model?
- Simple linear regression
  - ✓ Assessing model fits and assumptions
  - ✓ Model comparison: full vs reduced
- Categorical predictor
  - ✓ T-test
  - ✓ ANOVA
- Multiple linear regression
  - ✓ Categorical and continuous predictors (ANCOVA)
  - ✓ Standardizing and centering


Types of explanatory variables
Continuous
Categorical
Continuous AND categorical
1 variable
1 with 2 levels
GLM
1. Simple linear regression
2. T test
4. ANCOVA
2 or more
1 with 3 levels or more and/or many factors
GLM
5. Multiple linear regression
3. ANOVA

ANCOVA

You can have any number of categorical and/or continuous predictors, but as their number increases, the interpretation of results gets more complex.

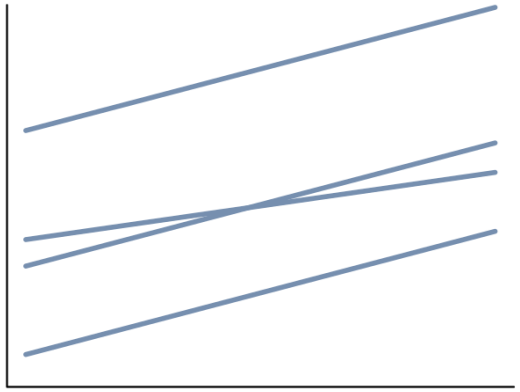
Frequently used ANCOVA models:

1. **One continuous and one categorical**
2. One continuous and two categorical
3. Two continuous and one categorical

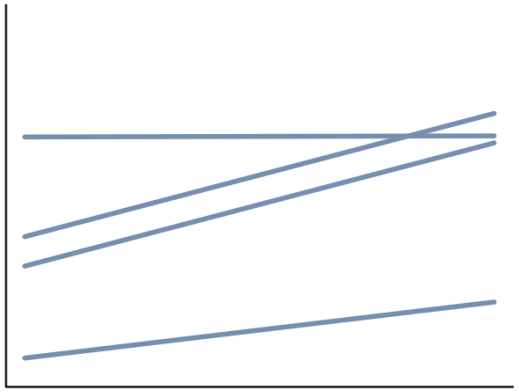
ANCOVA – one continuous and one categorical

- If only your categorical is significant, remove the continuous -> you will have a simple ANOVA
- If only your continuous is significant, remove your categorical -> you will have a simple linear regression
- If the interaction between your continuous and categorical (*) variables is significant, you should explore which levels of the factor have different slopes from the others.

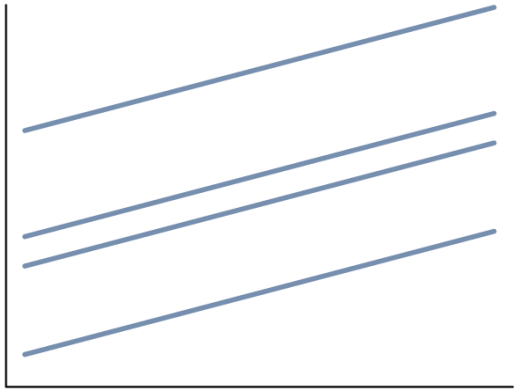
One level of the factor has a different slope



Many levels have different slopes



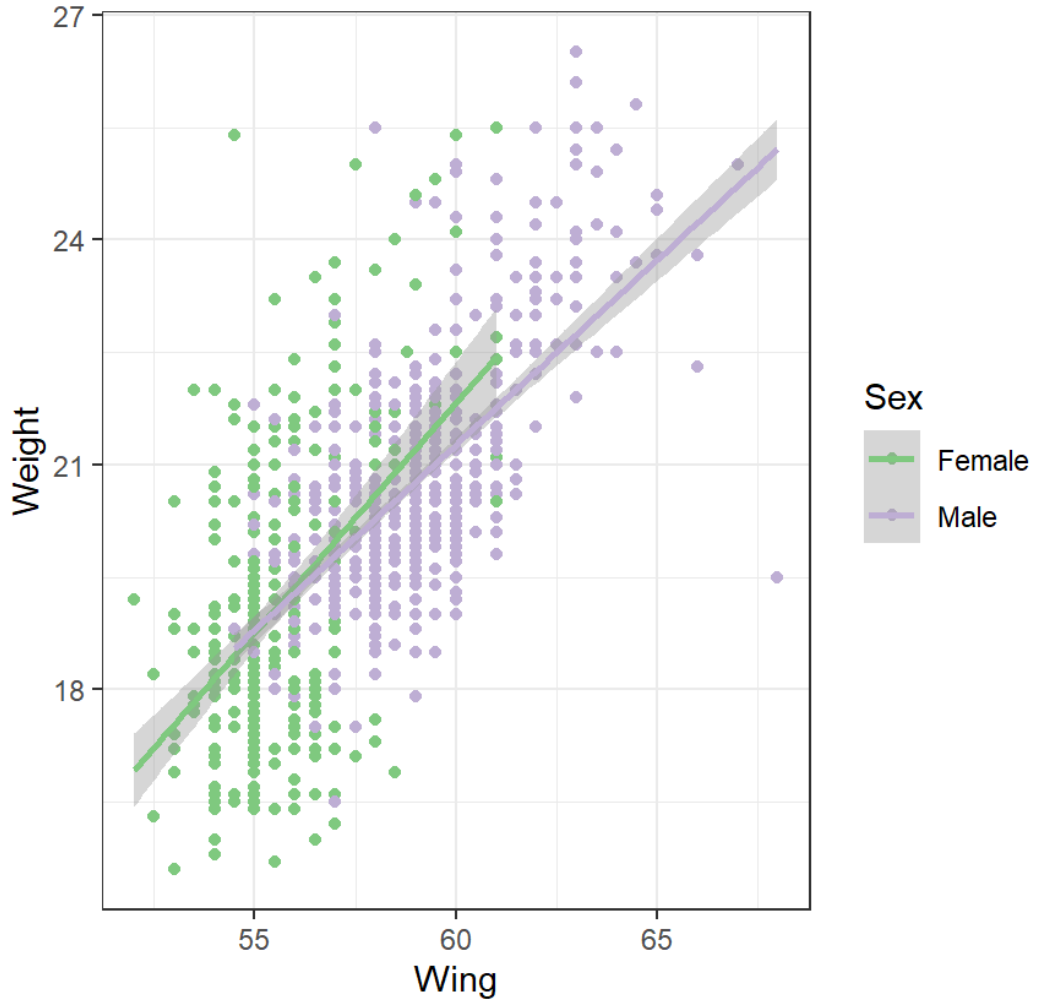
No interaction



ANCOVA – interaction between a continuous and categorical variable

Question: What is the relationship between wing length and weight for female and male Savannah sparrows?

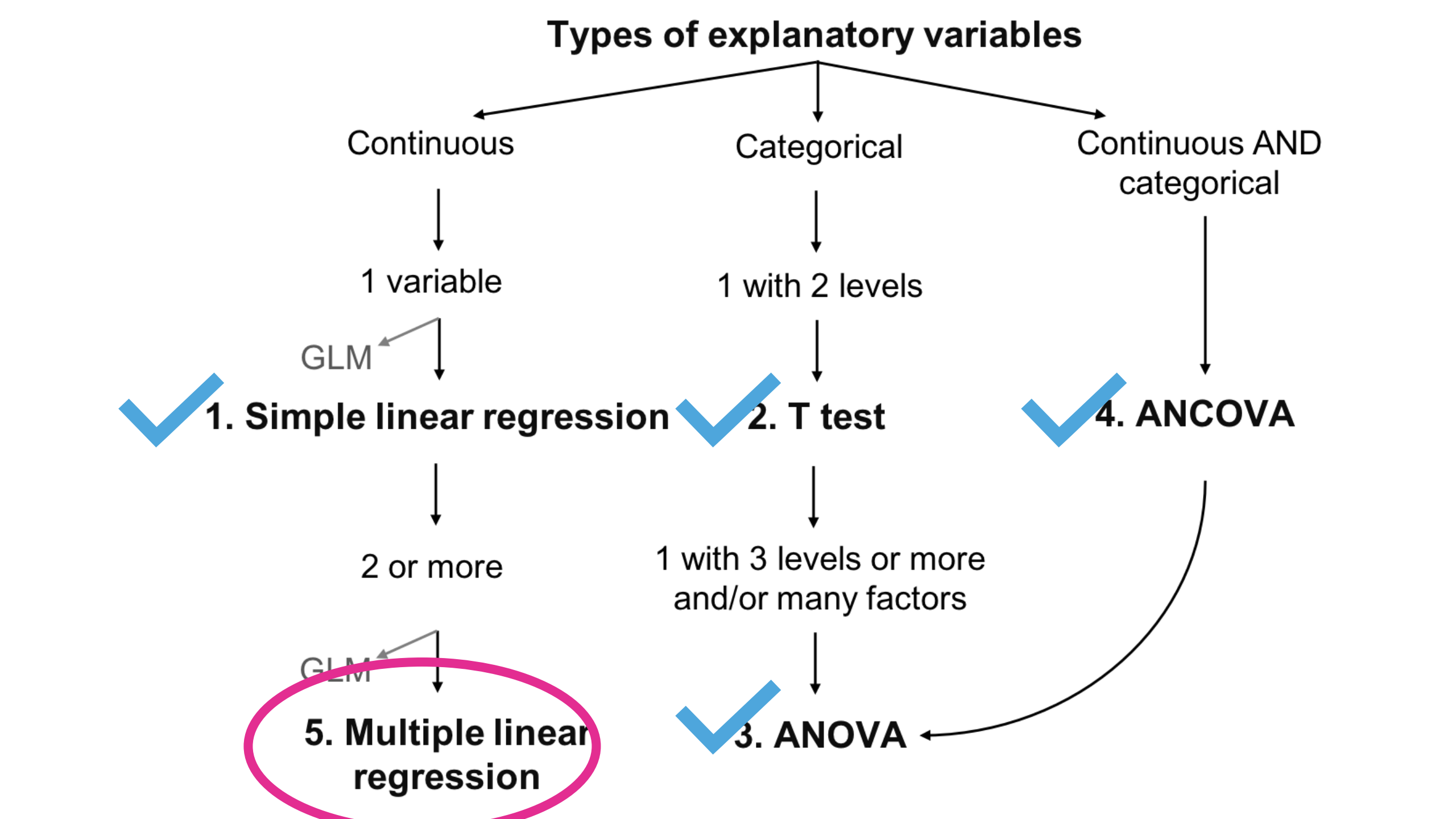
Hypothesis: For Savannah sparrows, the influence of **wing length (mm)** on **weight (grams)** will differ between males and females



ANCOVA – interaction between a continuous and categorical variable

```
> model4<-lm(data=spar, Weight~Wing*Sex)
> summary(model4)

Call:
lm(formula = Weight ~ Wing * Sex, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7137 -0.7649 -0.0772  0.7018  6.9476

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.95386    2.53165  -5.907 4.82e-09 ***
Wing          0.61296    0.04539  13.505  < 2e-16 ***
SexMale       6.59958    2.98288   2.212   0.0272 *
Wing:SexMale -0.11931    0.05273  -2.263   0.0239 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.326 on 975 degrees of freedom
Multiple R-squared:  0.4373,    Adjusted R-squared:  0.4355
F-statistic: 252.6 on 3 and 975 DF,  p-value: < 2.2e-16
```

```
> anova(model4)
Analysis of Variance Table

Response: Weight
           Df  Sum Sq Mean Sq  F value  Pr(>F)
Wing        1 1320.17 1320.17 750.9752 < 2e-16 ***
Sex         1    2.74    2.74   1.5609 0.21184
Wing:Sex    1    9.00    9.00   5.1201 0.02387 *
Residuals 975 1713.99    1.76
```


Types of explanatory variables
Continuous
Categorical
Continuous AND categorical
1 variable
1 with 2 levels
GLM
1. Simple linear regression
2. T test
4. ANCOVA
2 or more
1 with 3 levels or more and/or many factors
GLM
5. Multiple linear regression
3. ANOVA

Multiple linear regression:

Only difference to simple linear regression: **several predictor variables** are included in the model.

Variables

- y: Response variable (**continuous**)
- $x_1, x_2, \ldots, x_k$ Several predictor variables (**continuous** or **categorical**)

Assumed relationship

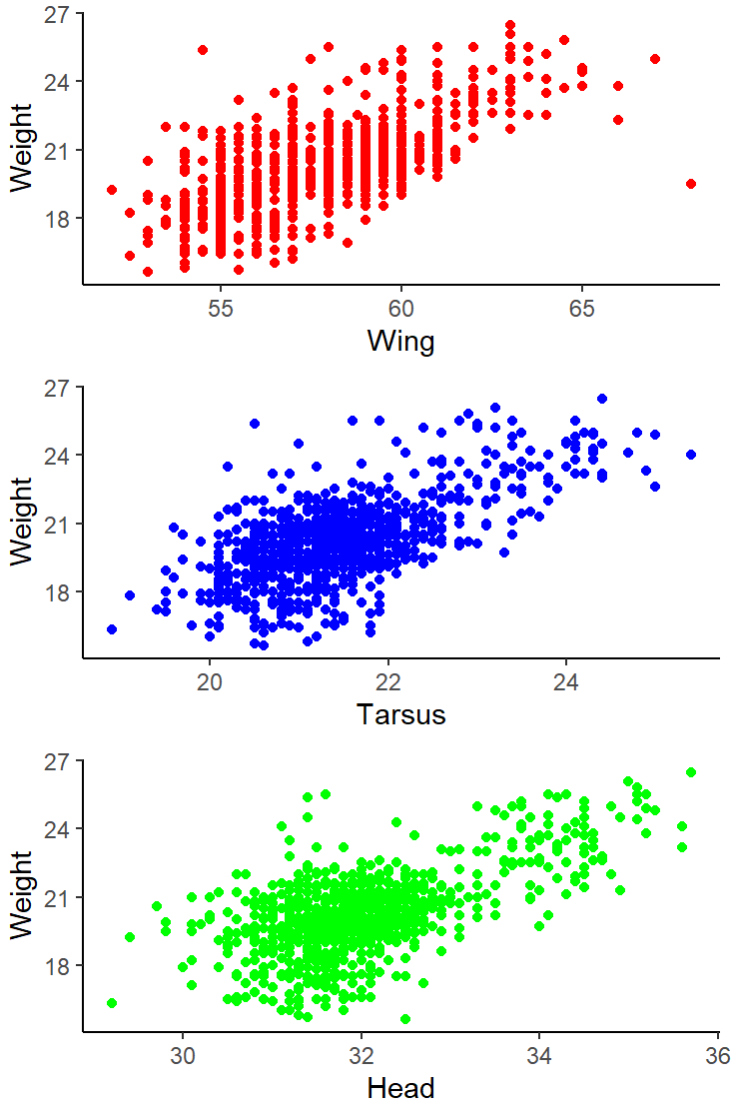
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_k x_{k,i} + \epsilon_i$$

- β_0 is the **intercept**
- $\beta_1, \beta_2, \ldots, \beta_k$ quantify the **effect** of $x_1, x_2, \ldots, x_k$ on y
- The residual ϵ_i captures **unexplained** variation
- The **fitted** (or predicted) value of y_i is defined as: $\hat{y}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_k x_{k,i}$

Multiple linear regression – no interaction

Question: Are wing length, tarsus length, and head size significant predictor of weight for Savannah sparrows?

Hypothesis: For Savannah sparrows, the **wing length (mm), tarsus length, and head size** of an individual influences the **weight (grams)** of an individual



Multiple linear regression

What is the strongest predictor variable?

```
> model4<-lm(data=spar, Weight~Wing+Tarsus+Head)
> summary(model4)

Call:
lm(formula = Weight ~ Wing + Tarsus + Head, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0298 -0.6723 -0.0770  0.6045  6.9316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.39915    1.29452 -16.531  < 2e-16 ***
Wing          0.30029    0.02036  14.748  < 2e-16 ***
Tarsus        0.48206    0.05915   8.150 1.11e-15 ***
Head          0.43374    0.05603   7.742 2.45e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.168 on 975 degrees of freedom
Multiple R-squared:  0.563,     Adjusted R-squared:  0.5617
F-statistic: 418.7 on 3 and 975 DF,  p-value: < 2.2e-16
```

Don’t trust these!

```
> anova(model4)
Analysis of Variance Table

Response: Weight
           Df  Sum Sq Mean Sq F value    Pr(>F)
Wing        1 1320.17 1320.17 967.016 < 2.2e-16 ***
Tarsus      1  312.85  312.85 229.159 < 2.2e-16 ***
Head        1   81.82   81.82  59.932  2.45e-14 ***
Residuals 975 1331.07    1.37
---
```

```
> model41<-lm(data=spar, Weight~Tarsus+Head+Wing)
> anova(model41)
Analysis of Variance Table

Response: Weight
           Df  Sum Sq Mean Sq F value    Pr(>F)
Tarsus      1 1227.52 1227.52  899.15 < 2.2e-16 ***
Head        1  190.38  190.38  139.45 < 2.2e-16 ***
Wing        1  296.94  296.94  217.51 < 2.2e-16 ***
Residuals 975 1331.07    1.37
```

Standardizing predictors

- Standardization refers to the process of subtracting the mean and dividing by the standard deviation
- Allows for direct comparison between coefficients in terms of effect size
- Helps when predictor variables are measured on very different scales – prevents predictors with larger scales from dominating
- Can mitigate issues related to collinearity (use with interactions)
- Does NOT change the test results
- Cannot communicate results in the original units of measurement

Multiple linear regression – standardization

```
> model4<-lm(data=spar, Weight~Wing+Tarsus+Head)
> summary(model4)

Call:
lm(formula = Weight ~ Wing + Tarsus + Head, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0298 -0.6723 -0.0770  0.6045  6.9316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.39915    1.29452 -16.531  < 2e-16 ***
Wing          0.30029    0.02036  14.748  < 2e-16 ***
Tarsus        0.48206    0.05915   8.150 1.11e-15 ***
Head          0.43374    0.05603   7.742 2.45e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.168 on 975 degrees of freedom
Multiple R-squared:  0.563,     Adjusted R-squared:  0.5617
F-statistic: 418.7 on 3 and 975 DF,  p-value: < 2.2e-16
```

```
> model5<-lm(data=spar, Weight~Wing_s+Tarsus_s+Head_s)
> summary(model5)

Call:
lm(formula = Weight ~ Wing_s + Tarsus_s + Head_s, data = spar)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0298 -0.6723 -0.0770  0.6045  6.9316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.22646    0.03734 541.643  < 2e-16 ***
Wing_s       0.68776    0.04663  14.748  < 2e-16 ***
Tarsus_s     0.44527    0.05464   8.150 1.11e-15 ***
Head_s       0.41561    0.05368   7.742 2.45e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.168 on 975 degrees of freedom
Multiple R-squared:  0.563,     Adjusted R-squared:  0.5617
F-statistic: 418.7 on 3 and 975 DF,  p-value: < 2.2e-16
```

Results remain the same, interpretation of coefficients changes

Extensions to linear models

What if the variance of the residuals is not constant & depends on another predictor?

- General least squares methods (gls) - Week 4

What if residuals are not independent because of nesting?

- Linear mixed effects models (lme) – Week 5

What if residuals are not independent because of autocorrelation or phylogeny?

- General least squares methods (gls) – Week 7

What if response data are binary or discrete?

- Generalized linear models (glm) – Week 8

What if the residuals are not independent and the response is binary or discrete?

- Generalized linear mixed models (glmm) – Week 11

Reminders

Thursday: Paper discussion and linear regression workshop

Office hours Monday and Thursday

Assignment #1 DUE Friday