Module 5: Linear regression and ANOVA

Jake Ferguson (jakeferg@hawaii.edu)

Module goals

1. Run simple- and multiple-regressions in R
2. Interpret regression outputs in R for continuous and categorical predictors
3. Run ANOVA in R

Sleep in mammals

Sleep is characterized by either slow wave (non-dreaming) or drapid eye movement (dreaming), with wide variability in theamount of both types of sleep.

- Roe Deer sleep < 3 hours/day
- Little Brown bat sleeps close to 20 hours per day

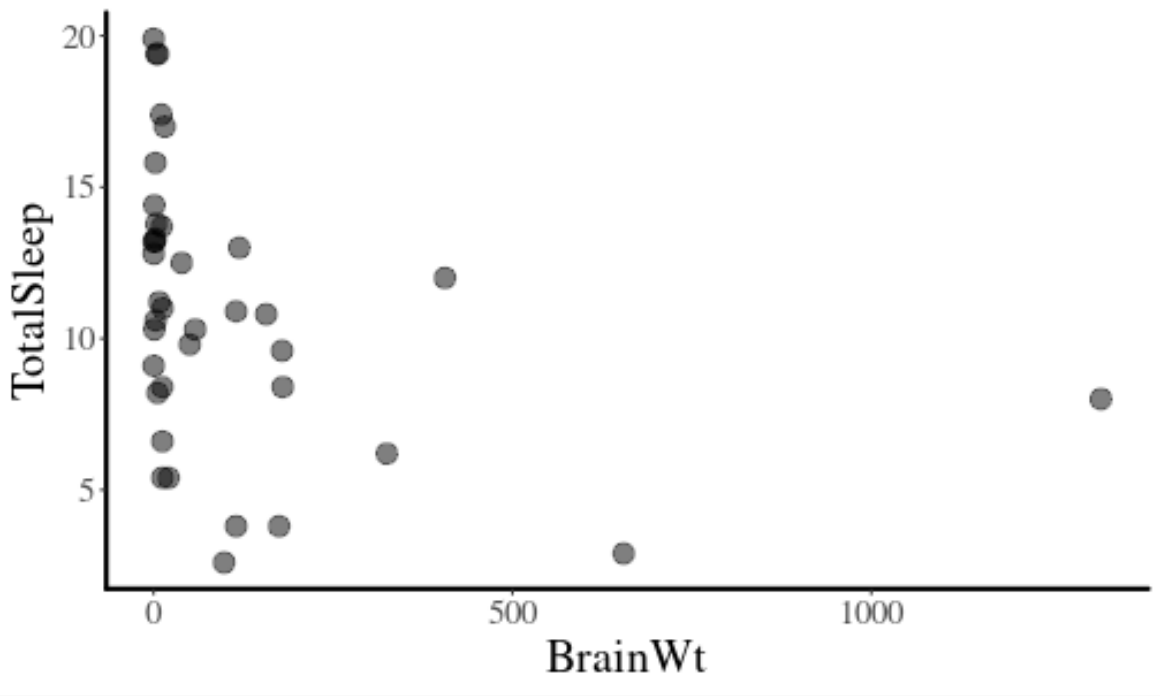
Why? What is the purpose of sleep?

- Memory consolidation
- Energy conservation

Allison, Truett and Cicchetti, Domenic V. (1976), Sleep in Mammals: Ecological and Constitutional Correlates, Science, November 12, vol. 194

Continuous response, continuous

Is the total amount of sleep related to an animals brain weight?



The linear correlation coefficient

Correlation is a measure of the strength and direction of linear association between two quantitative variables

Properties and assumptions

$$-1 \leq \rho \leq 1$$

$0 < \rho \leq 1$: positive association

$\rho = 1$: no association

$-1 \leq \rho < 0$: negative association

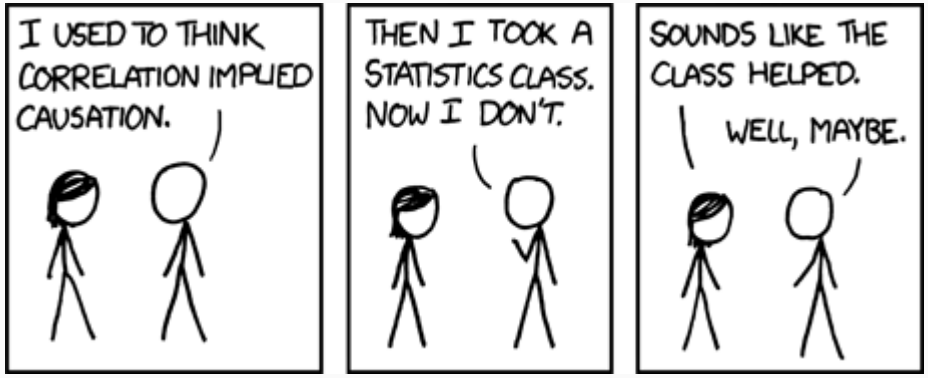
We denote the true correlation as ρ. The sample estimate of the correlation is r:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2)}\sqrt{\sum(Y_i - \bar{Y})^2)}}$$

- Our estimate, r, assumes random sampling and independent samples

Linear correlations and dependence

- $r = 0$ means no **linear** association. The variables could still be otherwise associated. Always plot your data!
- Recall definition of statistical independence: $P(Y|X) = P(Y)$
  - If $\rho \neq 0$ then X and Y are **definitely not** independent
  - If $\rho = 0$ then X and Y **may be** independent



https://www.xkcd.com

Getting pairwise correlations in R

```
cor(x=sleep.dat$BrainWt, y=sleep.dat$TotalSleep, use="pairwise.complete") #the argume
```

```
## [1] -0.33194
```

Correlations between multiple variables

The correlation matrix, $\boldsymbol{\Sigma}$, contains each pairwise correlation between $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$.

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{YX} & 1 & \rho_{YZ} \\ \rho_{ZX} & \rho_{ZY} & 1 \end{bmatrix}$$

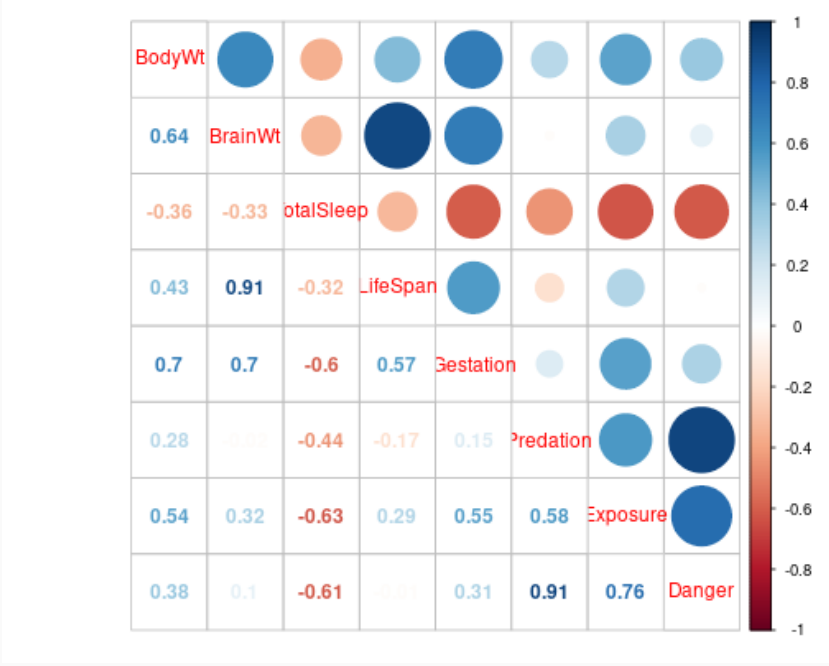
We are often interested primarily in correlations between predictor and the response, but we will discuss correlations between predictors as well.

Illustrating correlation matrices

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
M ← cor(sleep.dat[,3:10], use="pairwise.complete") #extract numeric responses
corrplot.mixed(M)
```

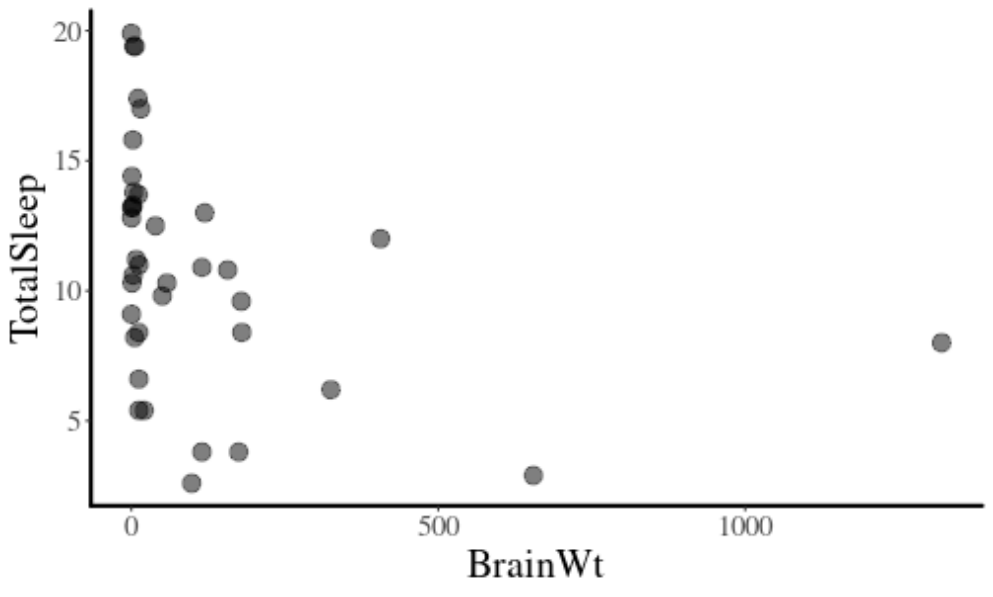


Exercise 5A

Calculating correlations

Limitations of correlation

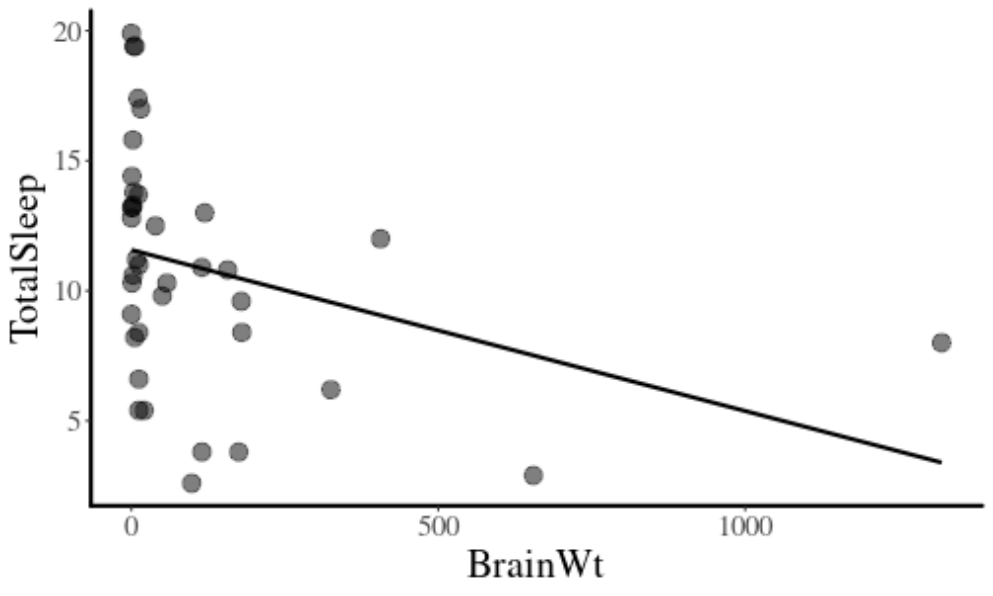
r tells us how strong a linear association is. It can't tell us how fast the response changes with changes in the predictor.



Simple linear regression

The goal is to fit a line that could have genererated the data. Estimates the **intercept** and slope of the line **slope**, as well as the **variance** in the residuals.

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$\varepsilon \sim \text{Norm}(0, \sigma^2)$$



Fitting works by trying to minimize σ^2.

Fitting `lm`'s in R

```
sleep.lm ← lm(TotalSleep ~ BrainWt, data=sleep.dat)
summary(sleep.lm)
```

```
##
## Call:
## lm(formula = TotalSleep ~ BrainWt, data = sleep.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3641 -2.9410 -0.1412  2.2388  8.3291
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.572422   0.770543  15.019   <2e-16 ***
## BrainWt     -0.006194   0.002934  -2.111   0.0418 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.322 on 36 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.1102,    Adjusted R-squared:  0.08547
## F-statistic: 4.458 on 1 and 36 DF,  p-value: 0.04175
```

Lets go through the output

- Slope and intercept estimates:

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 11.572422384 0.770543402 15.018521 4.450444e-17
## BrainWt     -0.006194472 0.002933888 -2.111352 4.175088e-02
```

- df (degrees of freedom used in the t-tests, df=n-1-#betas)

```
## [1] 36
```

- Residual standard error (σ, the standard deviation of the residuals)

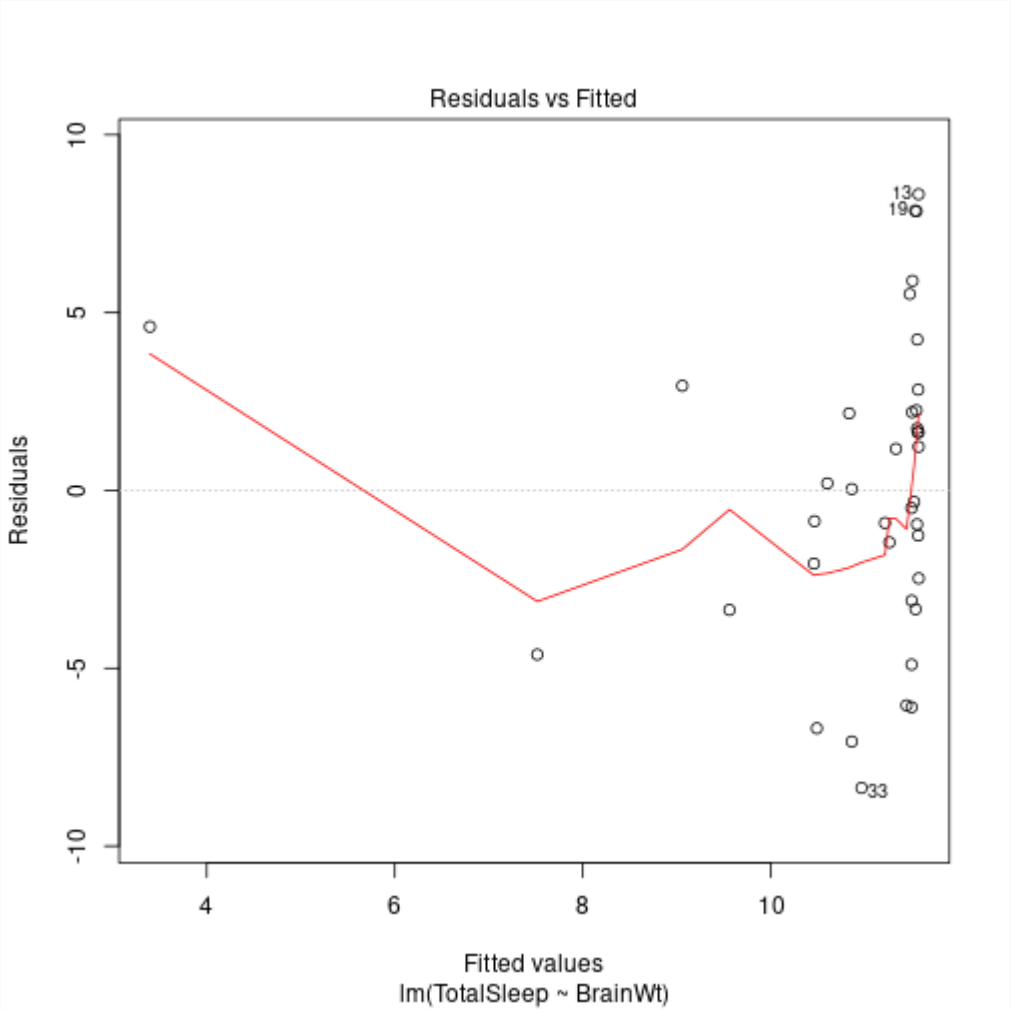
```
## [1] 4.321744
```

- Multiple R-squared (the squareroot of the correlation coefficient!)

```
## [1] 0.1101841
```

Diagnostic plots

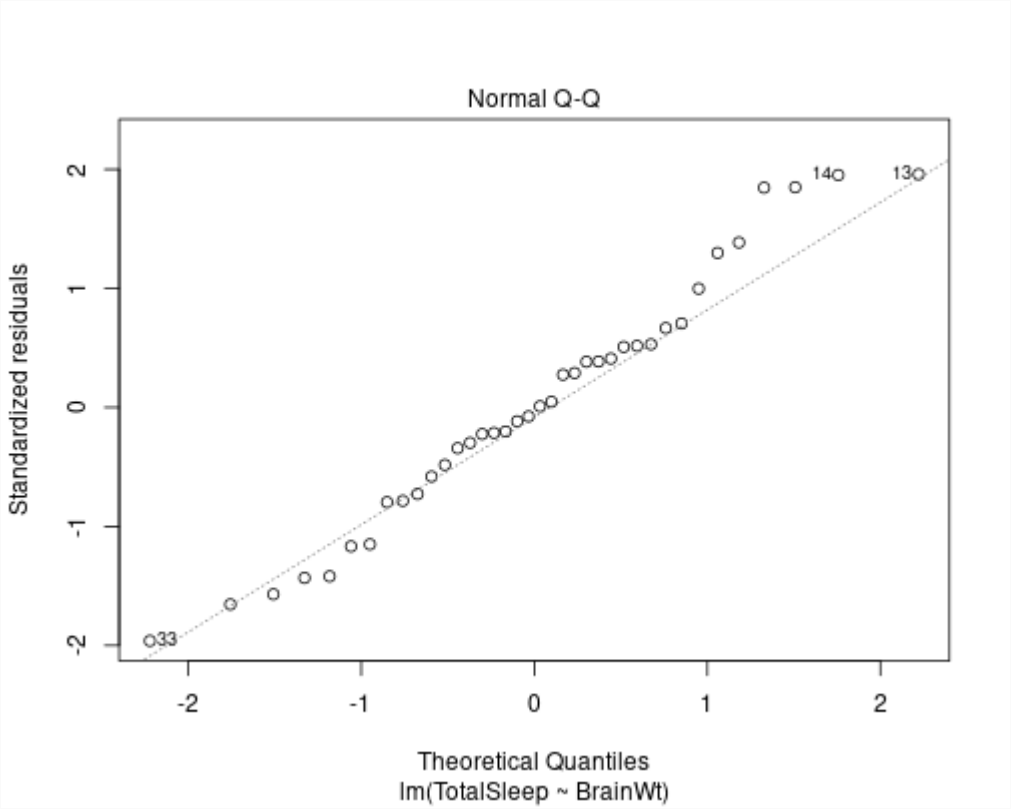
```
plot(sleep.lm, which=1)
```



The quantile-quantile plot

Do the residuals follow a normal distribution?

```
plot(sleep.lm, which=2)
```



Extracting confidence intervals

```
confint(sleep.lm)
```

```
##                    2.5 %        97.5 %
## (Intercept) 10.00968793 13.1351568339
## BrainWt     -0.01214467 -0.0002442709
```

Exercise 5B

Simple linear regression

Categorical predictors

Effects coding

```
fish.dat ← read.csv(file="../Data/Butte
effects.lm ← lm(Length ~ Sex, data=fish
round(summary(effects.lm)$coef[,c(1,2,4)
```

```
##             Estimate Std. Error Pr(>|t|)
## (Intercept)  93.6875     1.8228   0.0000
## SexMale       1.6250     2.5779   0.5332
```

```
head(model.matrix(effects.lm))
```

```
##   (Intercept) SexMale
## 1           1       1
## 2           1       1
## 3           1       1
## 4           1       1
## 5           1       1
## 6           1       1
```

Means coding

```
means.lm ← lm(Length ~ Sex-1, data=fish
round(summary(means.lm)$coef[,c(1,2,4)],
```

```
##            Estimate Std. Error Pr(>|t|)
## SexFemale  93.6875     1.8228        0
## SexMale    95.3125     1.8228        0
```

```
head(model.matrix(means.lm))
```

```
##   SexFemale SexMale
## 1         0       1
## 2         0       1
## 3         0       1
## 4         0       1
## 5         0       1
## 6         0       1
```

Effects coding

In the effects coding, the parameter estimate is the deviation from the intercept.

If $x = 0$, the effect is "off"

$$E[y] = \beta_0 + \beta_1 \cdot 0$$

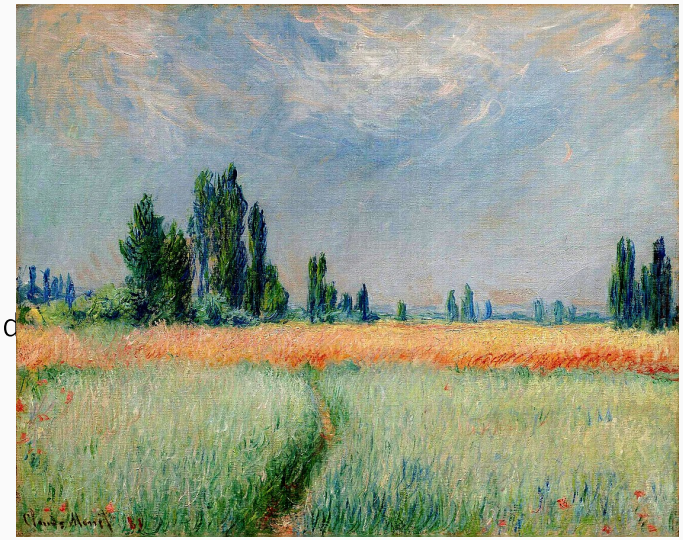
If $x = 1$, the effect is "on"

$$E[y] = \beta_0 + \beta_1 \cdot 1$$

Multiple regression

Prices of Monet paintings

```
monet.mod  ← lm(PRICE ~ SIZE + HOUSE, d
summary(monet.mod)
```



```
##
## Call:
## lm(formula = PRICE ~ SIZE + HOUSE, data = monet.d
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.273 -2.037 -1.012  0.662 27.126
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.3421516  0.6002300  -2.236   0.0259 *
## SIZE         0.0048434  0.0006022   8.043 9.59e-15 ***
## HOUSE2       0.5508707  0.4006872   1.375   0.1699
## HOUSE3      -0.4420160  0.8725283  -0.507   0.6127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Interpreting the output

```
summary(monet.mod)$coef
```

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -1.342151602 0.6002300295 -2.2360621 2.588637e-02
## SIZE         0.004843393 0.0006022223  8.0425343 9.594054e-15
## HOUSE2       0.550870700 0.4006872066  1.3748148 1.699416e-01
## HOUSE3      -0.442016004 0.8725282523 -0.5065922 6.127140e-01
```

Now the **intercept** parameter is the value when the `SIZE` of the painting is 0 and at the reference level (in this case auction house `HOUSE1`).

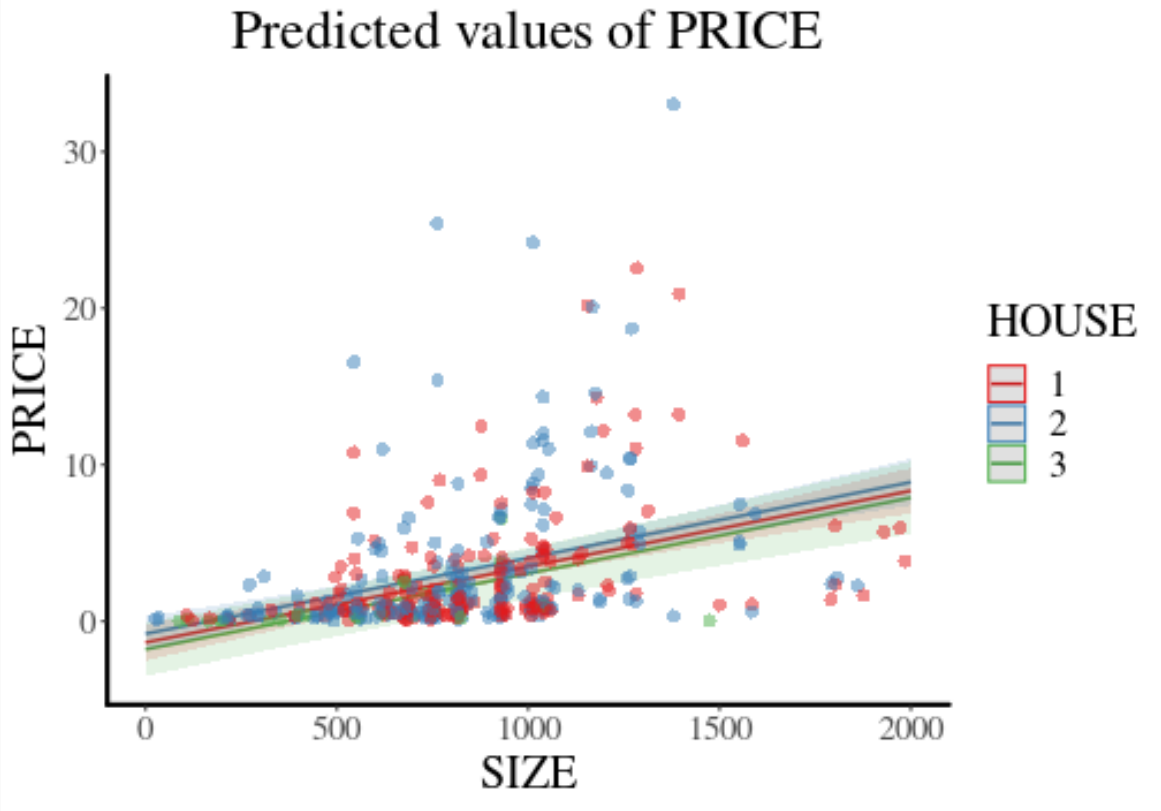
The `HOUSE2` parameter tells us that the average price is higher for second auction house relative to the first

The `HOUSE3` parameter tells us that the average price is lower for second auction house relative to the first

The **SLOPE** parameter of `SIZE` tells us how fast the price increases with the size of the painting.

Plotting the output

```
library(sjPlot)
plot_model(monet.mod, type="pred", terms=c("SIZE", "HOUSE"), show.data=TRUE)
```



Interactions

I the previous model the intercept varied by auction house. In this model **both** the intercept and slope vary.

```
monet.mod2  ← lm(PRICE ~ SIZE * HOUSE, data=monet.dat)
coef(summary(monet.mod2))
```

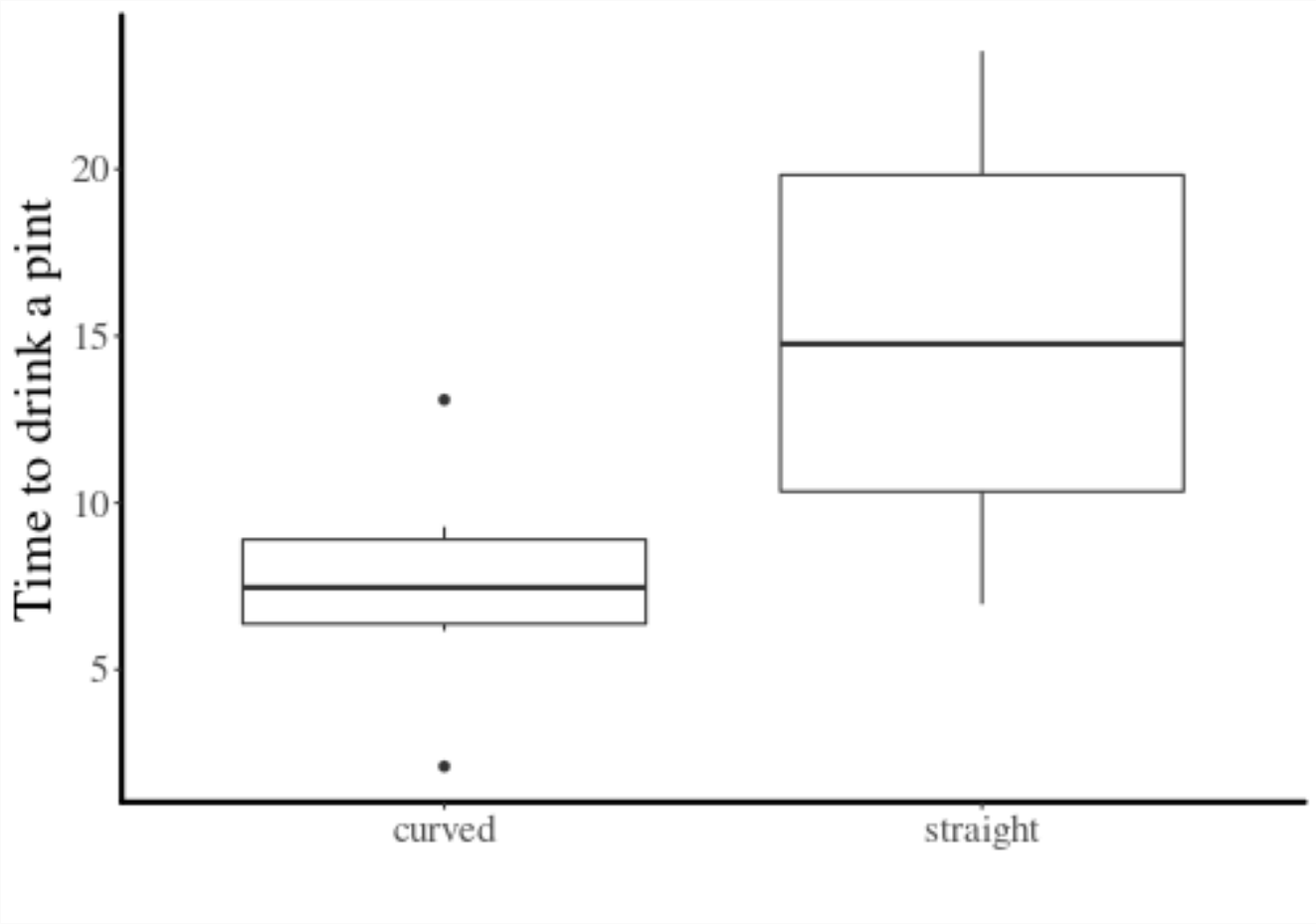
```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -1.1340749614 0.8288422480 -1.3682639 1.719843e-01
## SIZE         0.0046046089 0.0008904523  5.1710899 3.658083e-07
## HOUSE2      -0.1479775599 1.1259553086 -0.1314240 8.955048e-01
## HOUSE3       0.9710124172 1.7922670931  0.5417789 5.882673e-01
## SIZE:HOUSE2  0.0008386604 0.0012446358  0.6738199 5.008083e-01
## SIZE:HOUSE3 -0.0025995930 0.0025613783 -1.0149196 3.107471e-01
```

Exercise 5C

Multiple linear regression

Multiple discrete predictors

The t-test compared the mean between two groups

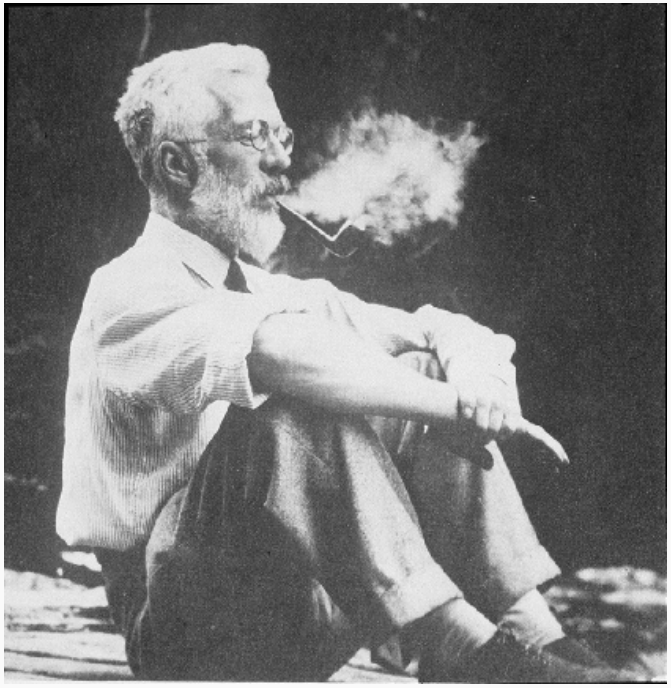


What about more than 2 groups?

image: https://www.homestratosphere.com

Testing multiple groups

- Do every pairwise comparison.
- Be clever.



http://www.economics.soton.ac.uk/staff/aldrich/fisherguide

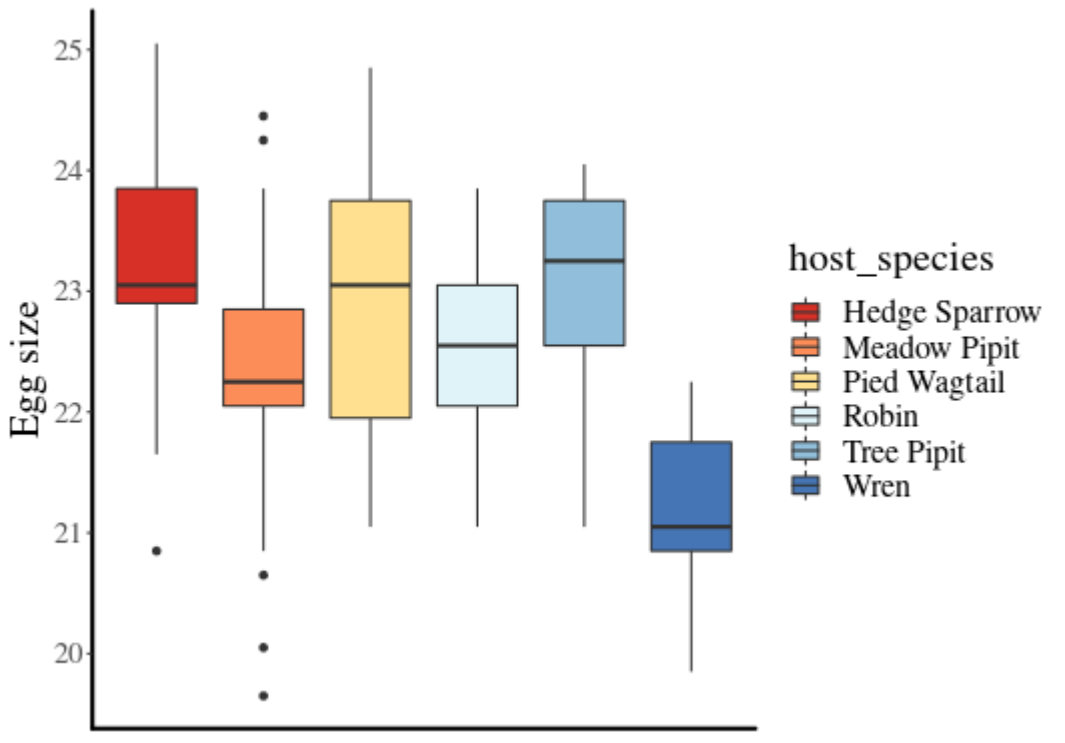
Example: Cuckoos

- Cuckoo birds lay their eggs in the nests of other birds
- When the cuckoo baby hatches, it kicks out all the original eggs/babies
- If the cuckoo is lucky, the mother will raise the cuckoo as if it were her own
- **Do cuckoo bird eggs found in nests of different species differ in size?**

In other words, does the typical egg size vary by the type of nest?

https://opinionator.blogs.nytimes.com/2010/06/01/cuckoo-cuckoo

The eye test



Intuition: If we remove a group from the population, does the overall variation in the data change?

The ANOVA model

$Y_1 \sim Norm(\mu_1, \sigma^2)$

$Y_2 \sim Norm(\mu_2, \sigma^2)$

$Y_3 \sim Norm(\mu_3, \sigma^2)$

$Y_4 \sim Norm(\mu_4, \sigma^2)$

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$?

assumption: same variance σ^2 for all groups!

Sums of squares

Variation between groups (SS_{groups}): $\sum_{j}^{G} n_j(\bar{Y}_j - \bar{Y})^2$.

distance from group mean to grand mean

$+$ Variation within groups (SS_{error}): $\sum_{j}^{G}\sum_{i}^{n_i}(Y_{i,j} - \bar{Y}_j)^2$.

distance from each observation to the group mean

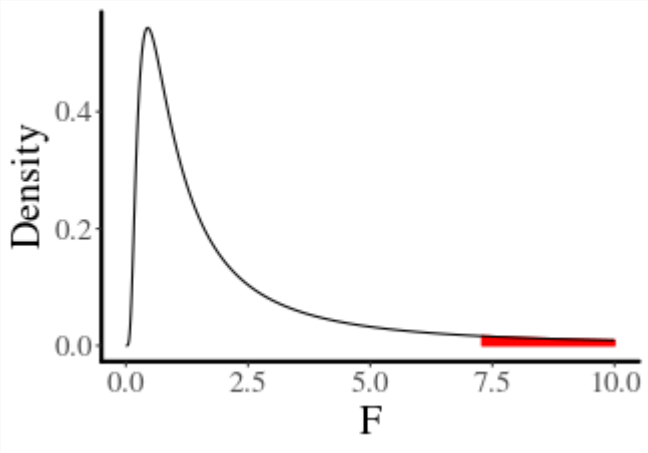
= Total variability : $\sum(Y - \bar{Y})^2$

If group means are consistent then we expect that the **variation between groups will be equal to the variation within groups**.

The F-test

$\mathbf{H_0}$: No differences among groups.

$$F = \frac{SS_{\text{groups}}/df_{groups}}{SS_{\text{error}}/df_{error}}$$



Now calculate the p-value.

Implementing ANOVA in R

```
summary(aov(egg_length ~ host_species, cuckoo.dat))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## host_species    5  42.94   8.588   10.39 3.15e-08 ***
## Residuals     114  94.25   0.827
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cool, there is a difference...

Exercise 5D

Run an ANOVA model

Post tests

Post (aka post-hoc) tests look at pairwise comparisons and attempt to control for the fact that you are making **lots of comparisons**.

The Tukey-test uses the t-statistic (mean/SE) but the p-value controls for the fact that you are making comparisons. It controls for 5% false positives for all tests, instead of per-test.

```
knitr::include_graphics("https://upload.wikimedia.org/wikipedia/en/e/e9/John_Tukey.jpg
```

Applying Tukey's test

```
TukeyHSD(aov(egg_length ~ host_species, cuckoo.dat))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = egg_length ~ host_species, data = cuckoo.dat)
##
## $host_species
##                                    diff          lwr         upr     p adj
## Meadow Pipit-Hedge Sparrow -0.82253968 -1.629133605 -0.01594576 0.0428621
## Pied Wagtail-Hedge Sparrow -0.21809524 -1.197559436  0.76136896 0.9872190
## Robin-Hedge Sparrow        -0.54642857 -1.511003196  0.41814605 0.5726153
## Tree Pipit-Hedge Sparrow   -0.03142857 -1.010892769  0.94803563 0.9999990
## Wren-Hedge Sparrow         -1.99142857 -2.970892769 -1.01196437 0.0000006
## Pied Wagtail-Meadow Pipit   0.60444444 -0.181375330  1.39026422 0.2324603
## Robin-Meadow Pipit          0.27611111 -0.491069969  1.04329219 0.9021876
## Tree Pipit-Meadow Pipit     0.79111111  0.005291337  1.57693089 0.0474619
## Wren-Meadow Pipit          -1.16888889 -1.954708663 -0.38306911 0.0004861
## Robin-Pied Wagtail         -0.32833333 -1.275604766  0.61893810 0.9155004
## Tree Pipit-Pied Wagtail     0.18666667 -0.775762072  1.14909541 0.9932186
## Wren-Pied Wagtail          -1.77333333 -2.735762072 -0.81090459 0.0000070
## Tree Pipit-Robin            0.51500000 -0.432271433  1.46227143 0.6159630
## Wren-Robin                 -1.44500000 -2.392271433 -0.49772857 0.0003183
```

image: https://en.wikipedia.org/wiki/John_Tukey

good wikipedia article: https://en.wikipedia.org/wiki/Tukey%27s_range_test

A word about anova

With multiple predictors, `aov` performs “sequential” tests (where order of entry matters!)

These tests are usually less appropriate than the t-tests from the summary function, especially for unbalanced designs.

A more robust approach uses type II ANOVA. This is implemented as the function `Anova` in the `car` library (helpful with categorical variables with multiple levels).

Even better is a randomization approach such as PERMANOVA in the `vegan` package

Summary

- Fit continuous response data to one or more predictors with regression
- Interpret output of model fits and hypothesis tests
- Introduction to diagnostic figures