Day 2 - File 03 - Statistical Tests

Melinda Higgins

7/29/2020

Let's look at some more plots

Let's look at some more plots of variables we think may be related to predicting the age of the abalones.

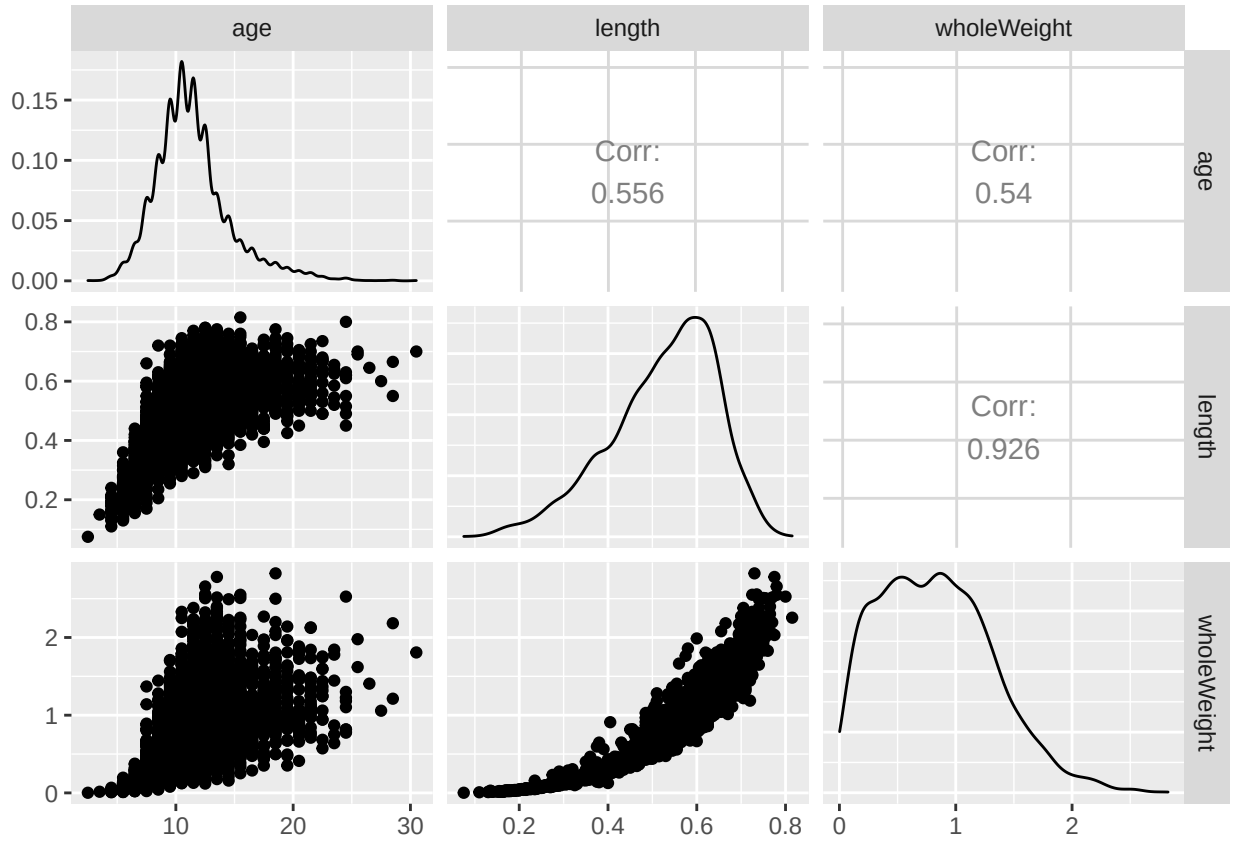
Let's look at age with length and whole weight.

For this let's try to **GGally** package which makes some cool plots especially matrix scatterplots. It is an extension package for **ggplot2**.

Learn more about **GGally** package at https://cran.r-project.org/web/packages/GGally/index.html.

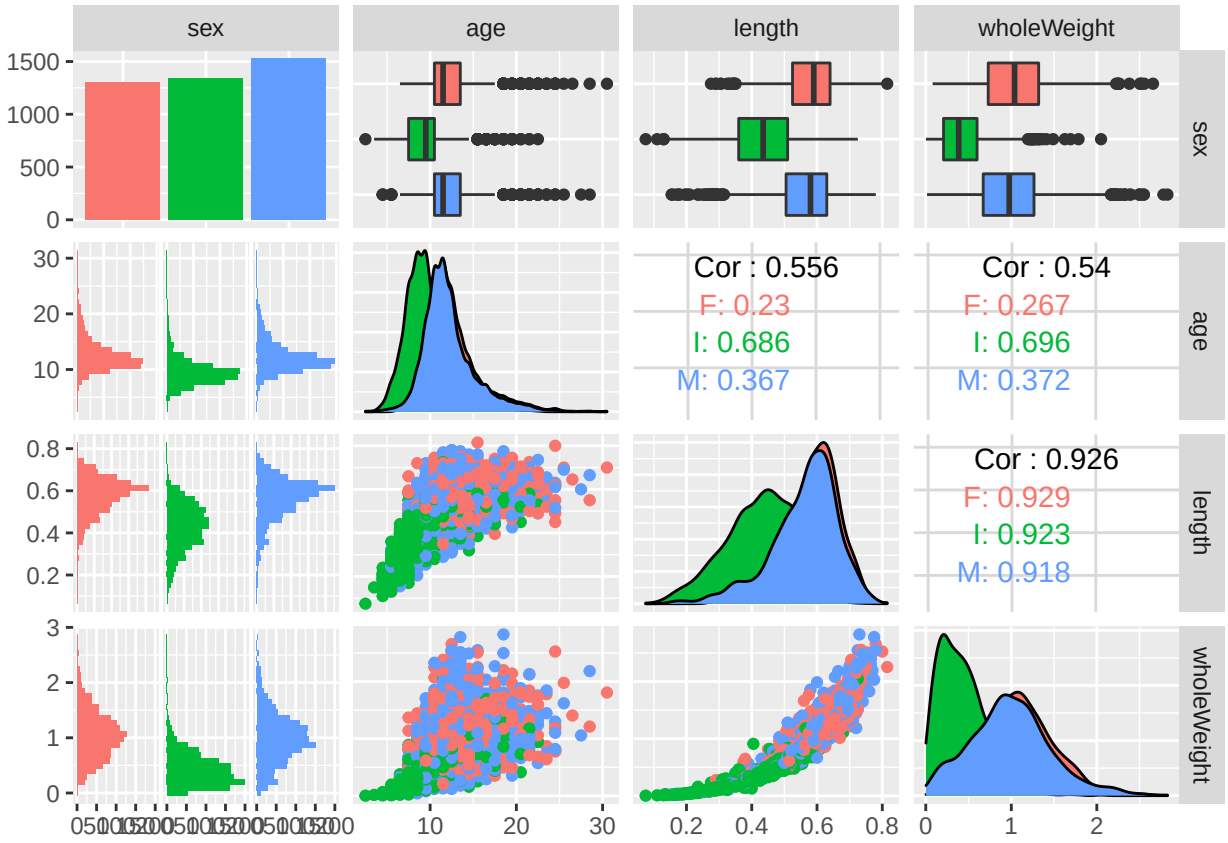
Also check out the "R Graph Gallery" for ideas, see https://www.r-graph-gallery.com/199-correlation-matrix-with-ggally.html.

```
library(GGally)
abaloneMod %>%
  select(age, length, wholeWeight) %>%
  GGally::ggpairs()
```



Color the points by sex - add this aesthetic to the `ggpairs()` function.

```
abaloneMod %>%
  select(sex, age, length, wholeWeight) %>%
  GGally::ggpairs(aes(color=sex))
```



Fit linear models

Since I'd like to compare models, I need to work with a dataset that has no missing data. We can create a dataset with no missing values across all variables as follows.

```
abaloneMod_complete <- abaloneMod %>%
  filter(complete.cases(abaloneMod))
```

Fit 3 models and save the results and look at the summary of each model.

Model 1 age by length.

```
lm1 <- lm(age ~ length, data = abaloneMod_complete)
lm1
```

```
##
## Call:
## lm(formula = age ~ length, data = abaloneMod_complete)
##
## Coefficients:
## (Intercept)       length
##        3.61        14.93
```

```
slm1 <- summary(lm1)
slm1
```

```
##
## Call:
## lm(formula = age ~ length, data = abaloneMod_complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9664 -1.6970 -0.7424  0.8842 16.6763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.6100     0.1863   19.38   <2e-16 ***
## length       14.9339     0.3463   43.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.681 on 4167 degrees of freedom
## Multiple R-squared:  0.3086, Adjusted R-squared:  0.3084
## F-statistic:  1860 on 1 and 4167 DF,  p-value: < 2.2e-16
```

Model 2 age by wholeWeight

```
lm2 <- lm(age ~ wholeWeight, data = abaloneMod_complete)
slm2 <- summary(lm2)
slm2
```

```
##
## Call:
## lm(formula = age ~ wholeWeight, data = abaloneMod_complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2621 -1.7529 -0.6926  1.0184 15.7026
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.49752    0.08268  102.78   <2e-16 ***
## wholeWeight  3.54628    0.08579   41.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.715 on 4167 degrees of freedom
## Multiple R-squared:  0.2908, Adjusted R-squared:  0.2907
## F-statistic:  1709 on 1 and 4167 DF,  p-value: < 2.2e-16
```

Model 3 age by length and wholeWeight

```
lm3 <- lm(age ~ length + wholeWeight,
          data=abaloneMod_complete)
slm3 <- summary(lm3)
slm3
```

```
##
## Call:
## lm(formula = age ~ length + wholeWeight, data = abaloneMod_complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9966 -1.6784 -0.7441  0.9198 16.3471
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9410     0.3176  15.555  < 2e-16 ***
## length       10.5709     0.9126  11.583  < 2e-16 ***
## wholeWeight   1.1528     0.2232   5.165 2.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.672 on 4166 degrees of freedom
## Multiple R-squared:  0.313,  Adjusted R-squared:  0.3126
## F-statistic: 948.8 on 2 and 4166 DF,  p-value: < 2.2e-16
```

Compare the full model `lm3` to the reduced models `lm1` or `lm2`.

```
anova(lm1, lm3)
```

```
## Analysis of Variance Table
##
## Model 1: age ~ length
## Model 2: age ~ length + wholeWeight
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   4167 29944
## 2   4166 29753  1     190.5 26.673 2.522e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm2, lm3)
```

```
## Analysis of Variance Table
##
## Model 1: age ~ wholeWeight
## Model 2: age ~ length + wholeWeight
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   4167 30711
## 2   4166 29753  1    958.22 134.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Make tables of the coefficients of each model

```
library(knitr)
knitr::kable(slm1$coefficients,
             caption = "Model 1 Age by Length")
```

Table 1: Model 1 Age by Length

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.610039	0.1862663	19.38107	0
length	14.933905	0.3463122	43.12266	0

```
knitr::kable(slm2$coefficients,
             caption = "Model 2 Age by Whole Weight")
```

Table 2: Model 2 Age by Whole Weight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.497525	0.0826756	102.78152	0
wholeWeight	3.546281	0.0857854	41.33898	0

```
knitr::kable(slm3$coefficients,
             caption = "Model 3 Age by Length and Whole Weight")
```

Table 3: Model 3 Age by Length and Whole Weight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.941020	0.3176445	15.555186	0e+00
length	10.570916	0.9126122	11.583140	0e+00
wholeWeight	1.152848	0.2232207	5.164609	3e-07

Another way to compare models

Use the **stargazer** package to compare models

The default output type is "latex" for making PDF documents. You will need to change this option depending on if you are knitting to HTML or to PDF. This will NOT work for DOC files.

The code chunk option must be set to `results='asis'`.

```
library(stargazer)

# set for HTML output
# stargazer(lm1, lm2, lm3, type="html")

# uncomment this if knitting to PDF
stargazer(lm1, lm2, lm3, type="latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jul 29, 2020 - 10:27:47 PM

Table 4:

	Dependent variable:		
	age		
	(1)	(2)	(3)
length	14.934***		10.571***
	(0.346)		(0.913)
wholeWeight		3.546***	1.153***
		(0.086)	(0.223)
Constant	3.610***	8.498***	4.941***
	(0.186)	(0.083)	(0.318)
Observations	4,169	4,169	4,169
R^2	0.309	0.291	0.313
Adjusted R^2	0.308	0.291	0.313
Residual Std. Error	2.681 (df = 4167)	2.715 (df = 4167)	2.672 (df = 4166)
F Statistic	1,859.564*** (df = 1; 4167)	1,708.911*** (df = 1; 4167)	948.847*** (df = 2; 4166)
Note:			*p<0.1; **p<0.05; ***p<0.01