Abalone Report

Data exploration, Cleaning, Models and Tests

Melinda Higgins

09/06/2023

Summary Statistics of Abalones' - Dimensional Measurements

A useful R package for making tables is the `arsenal` package. Learn more at:

- https://cran.r-project.org/web/packages/arsenal/index.html and
- https://mayoverse.github.io/arsenal/.

The key function is `tableby()`, see the vignette at https://mayoverse.github.io/arsenal/articles/tableby.html.

	Overall (N=4169)
length	
Mean (SD)	0.524 (0.120)
Range	0.075 - 0.815
diameter	
Mean (SD)	0.408 (0.099)
Range	0.055 - 0.650
height	
Mean (SD)	0.139 (0.039)
Range	0.010 - 0.515

Summary Statistics of Abalones' - Weight Measurements

	Overall (N=4169)
wholeWeight	
Mean (SD)	0.830 (0.490)
Range	0.002 - 2.825
shuckedWeight	
Mean (SD)	0.360 (0.222)
Range	0.001 - 1.488
visceraWeight	
Mean (SD)	0.181 (0.110)
Range	0.000 - 0.760
shellWeight	
Mean (SD)	0.239 (0.139)
Range	0.002 - 1.005

Abalone Dimensional Measurements by Sex - default statistical tests

Now we can add a grouping variable such as comparing these summary statistics between the 3 biological sex groups: Male, Female and Infant.

Notice that the default settings produce a p-value. The `arsenal::tableby()` function is performing an ANOVA (analysis of variance) for each of these measurements.

	F (N=1306)	I (N=1335)	M (N=1528)	Total (N=4169)	p value
length					< 0.001[1]
Mean (SD)	0.579 (0.086)	0.428 (0.109)	0.561 (0.103)	0.524 (0.120)	
Range	0.275 - 0.815	0.075 - 0.725	0.155 - 0.780	0.075 - 0.815	
diameter					< 0.001[1]
Mean (SD)	0.455 (0.071)	0.327 (0.088)	0.439 (0.084)	0.408 (0.099)	
Range	0.195 - 0.650	0.055 - 0.550	0.110 - 0.630	0.055 - 0.650	
height					< 0.001[1]
Mean (SD)	0.157 (0.030)	0.108 (0.032)	0.151 (0.035)	0.139 (0.039)	
Range	0.015 - 0.250	0.010 - 0.220	0.025 - 0.515	0.010 - 0.515	

1. Linear Model ANOVA

Abalone Dimensional Measurements by Sex - change statistical test

Suppose you decide that `diameter` is skewed and really need non-parametric statistics and the Kruskall-Wallis non-parametric ANOVA test performed. We can customize the statistics - learn more at https://mayoverse.github.io/arsenal/articles/tableby.html#change-summary-statistics-within-the-formula-1 and see more options at https://mayoverse.github.io/arsenal/articles/tableby.html#available-function-options-1.

You'll notice this automatically creates footnotes for each customized statistic in the output table.

	F (N=1306)	I (N=1335)	M (N=1528)	Total (N=4169)	p value
length					< 0.001[1]
Mean (SD)	0.579 (0.086)	0.428 (0.109)	0.561 (0.103)	0.524 (0.120)	
diameter					< 0.001[2]
Median (Q1, Q3)	0.465 (0.410, 0.505)	0.335 (0.270, 0.395)	0.455 (0.395, 0.500)	0.425 (0.350, 0.480)	
height					< 0.001[1]
Mean (SD)	0.157 (0.030)	0.108 (0.032)	0.151 (0.035)	0.139 (0.039)	

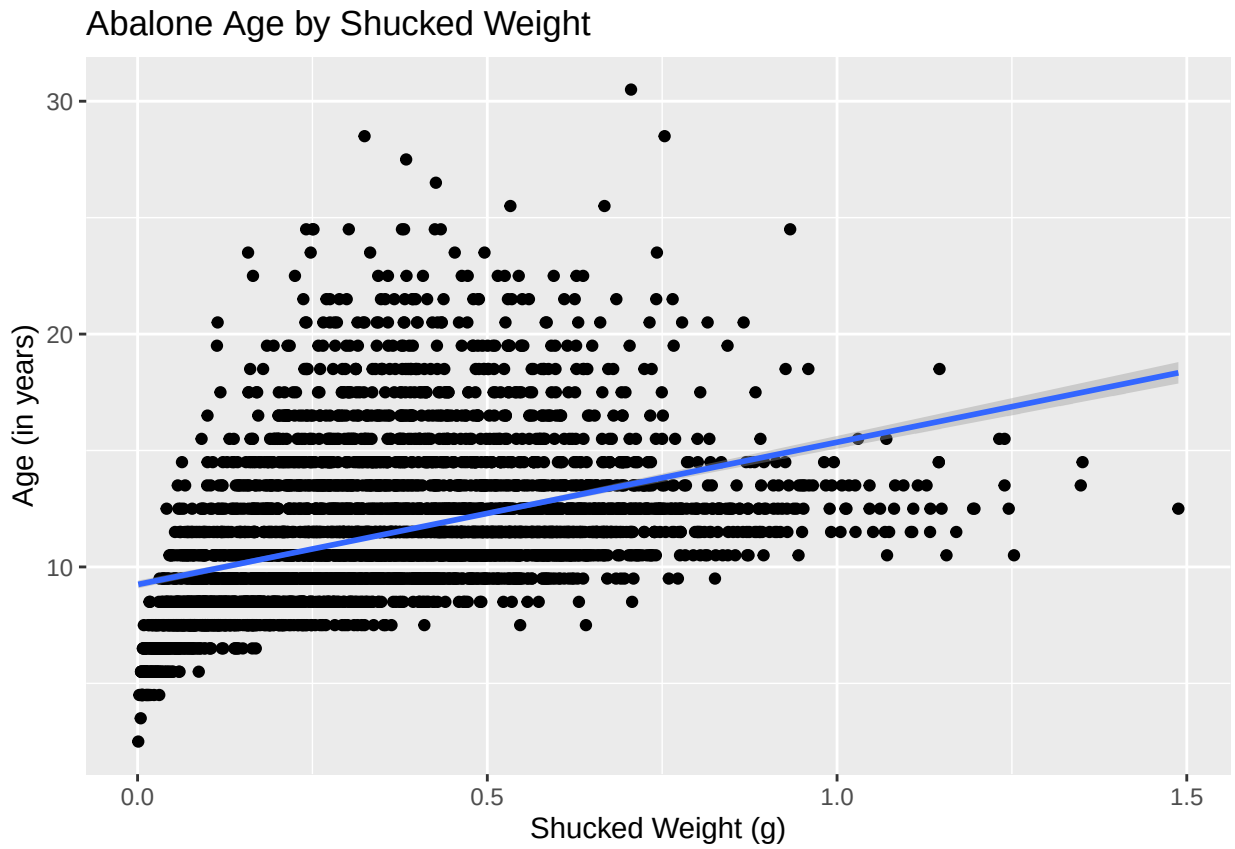
1. Linear Model ANOVA
2. Kruskal-Wallis rank sum test

Abalone Weight Measurements by Sex

	F (N=1306)	I (N=1335)	M (N=1528)	Total (N=4169)	p value
wholeWeight					$< 0.001^1$
Mean (SD)	1.047 (0.430)	0.432 (0.286)	0.991 (0.471)	0.830 (0.490)	
Range	0.080 - 2.657	0.002 - 2.050	0.015 - 2.825	0.002 - 2.825	
shuckedWeight					$< 0.001^1$
Mean (SD)	0.446 (0.199)	0.191 (0.128)	0.433 (0.223)	0.360 (0.222)	
Range	0.031 - 1.488	0.001 - 0.773	0.006 - 1.351	0.001 - 1.488	
visceraWeight					$< 0.001^1$
Mean (SD)	0.231 (0.098)	0.092 (0.063)	0.216 (0.105)	0.181 (0.110)	
Range	0.021 - 0.590	0.000 - 0.440	0.003 - 0.760	0.000 - 0.760	
shellWeight					$< 0.001^1$
Mean (SD)	0.302 (0.126)	0.128 (0.085)	0.282 (0.131)	0.239 (0.139)	
Range	0.025 - 1.005	0.002 - 0.655	0.005 - 0.897	0.002 - 1.005	

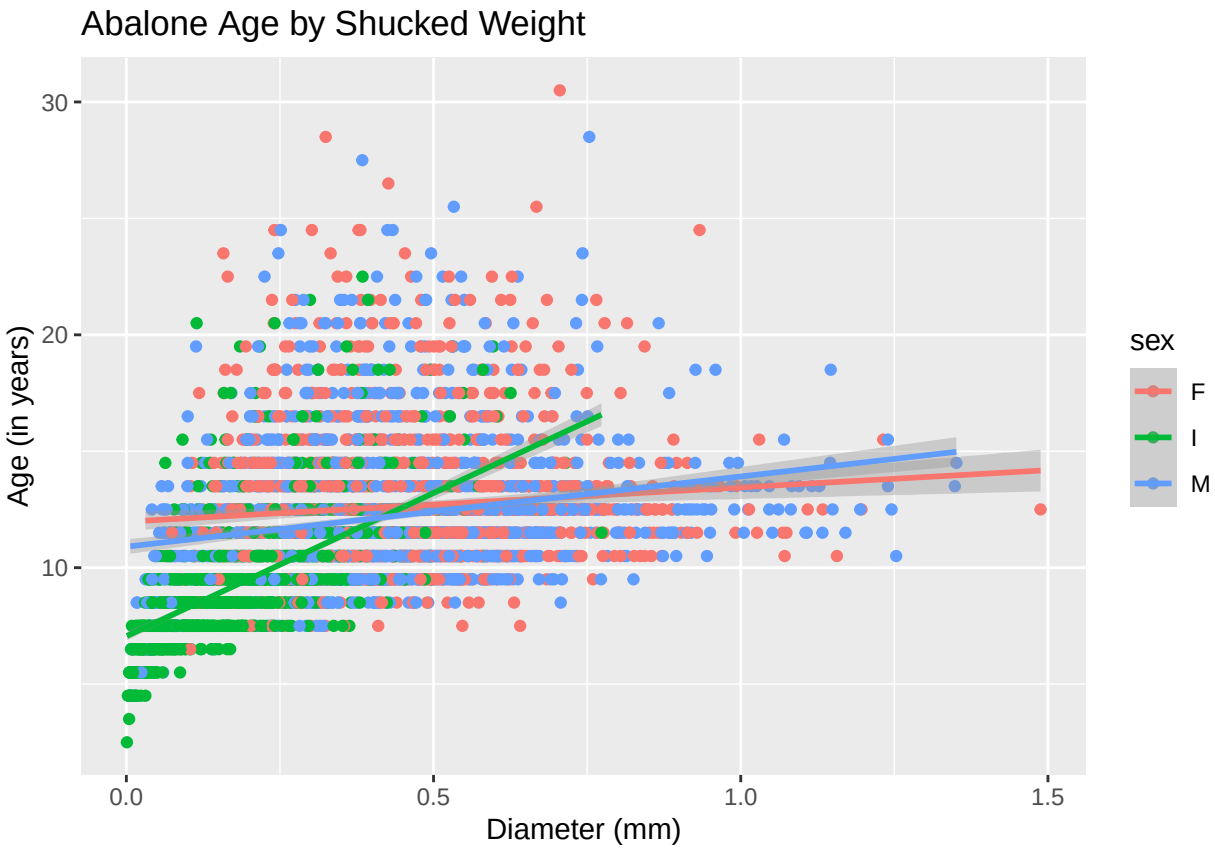
1. Linear Model ANOVA

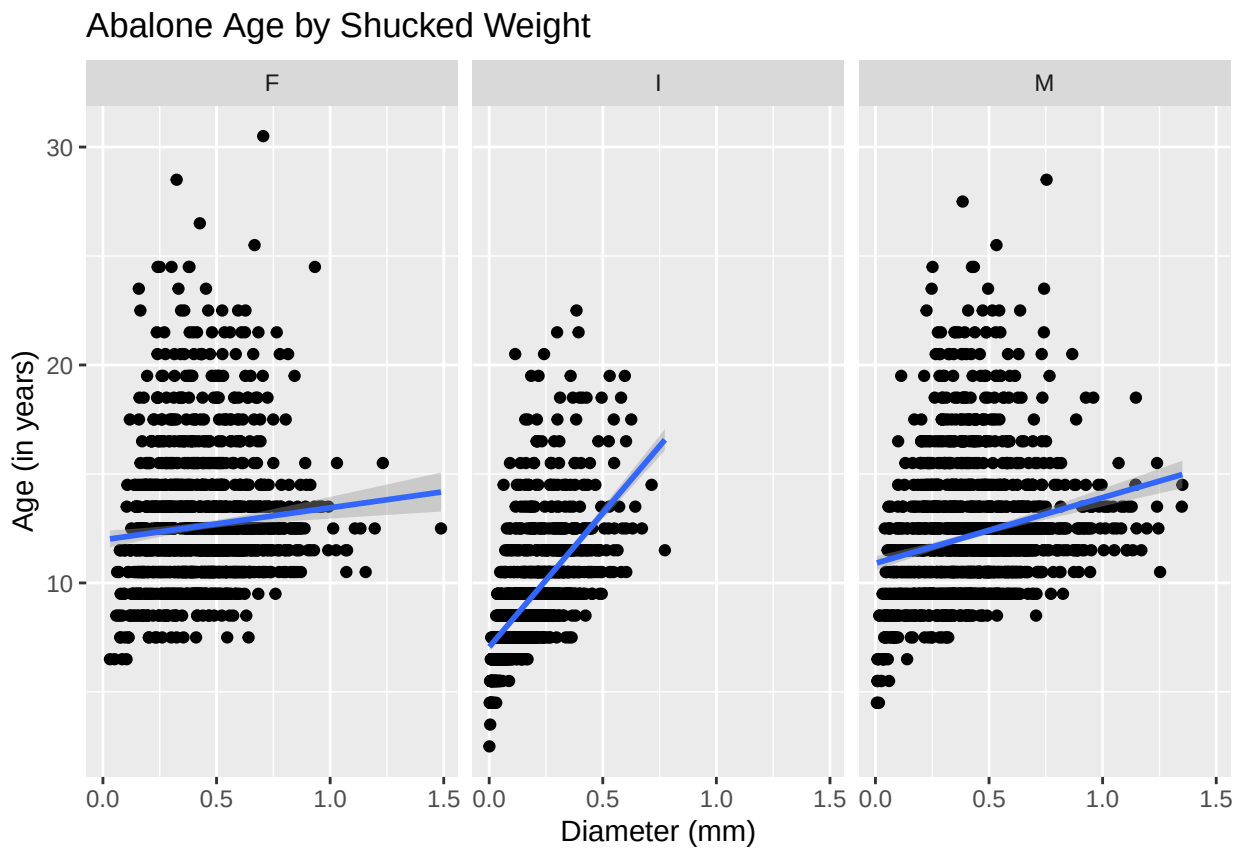
Plot of Abalone Age by shuckedWeight



Plot of Abalone Age by Shucked Weight - by sex

Create a plot of abalone age by shucked weight in g Show the plot by sex - either add a color by sex or a facet_wrap().


Abalone Age by Shucked Weight
Age (in years)
30
20
10
0.0
0.5
1.0
1.5
Diameter (mm)
sex
F
I
M



Linear Regression - Abalone Age by Shucked Weight (model 1)

Table 6: Regression of Abalone Age by Shucked Weight

term	estimate	std.error	statistic	p.value
(Intercept)	9.243295	0.0861858	107.24850	0
shuckedWeight	6.110826	0.2039592	29.96102	0

Linear Regression - Abalone Age by Shucked Weight, adjusted for sex (model 2)

Table 7: Regression of Abalone Age by Shucked Weight adjusted for Sex

term	estimate	std.error	statistic	p.value
(Intercept)	10.9143080	0.1285277	84.917946	0.0000000
shuckedWeight	3.8482792	0.2295312	16.765820	0.0000000
sexI	-2.2489742	0.1239592	-18.142855	0.0000000
sexM	-0.3749078	0.1057687	-3.544601	0.0003975

Compare Models - piecemeal steps

The change in R2 for the 2 models is 0.0657824 with a p-value of $4.3055706 \times 10^{-76}$.

Compare models - use `gtsummary` package

Model 1 output

Characteristic	Beta	95% CI	p-value
(Intercept)	9.2	9.1, 9.4	<0.001
shuckedWeight	6.1	5.7, 6.5	<0.001

Model 2 output

Characteristic	Beta	95% CI	p-value
(Intercept)	11	11, 11	<0.001
shuckedWeight	3.8	3.4, 4.3	<0.001
sex			
F	—	—	
I	-2.2	-2.5, -2.0	<0.001
M	-0.37	-0.58, -0.17	<0.001

Put models side by side

Characteristic	Beta	95% CI	p-value	Beta	95% CI	p-value
(Intercept)	9.2	9.1, 9.4	<0.001	11	11, 11	<0.001
shuckedWeight	6.1	5.7, 6.5	<0.001	3.8	3.4, 4.3	<0.001
sex						
F				—	—	
I				-2.2	-2.5, -2.0	<0.001
M				-0.37	-0.58, -0.17	<0.001

The `stargazer` package - works for HTML and PDF

WARNING This does NOT work for WORD documents. However, you can create the HTML output and then "cut-and-paste" the HTML table into WORD.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Wed, Sep 06, 2023 - 9:54:43 PM

Creating APA style tables

This package makes nice output but saves each table in a separate output DOC file. This does work for WORD documents.

To embed the `apaTable` output inside the Rmarkdown document, we can do the following...

Pull out the key parts of the output object. Make a nice table for the formatted output for `apa3` and then add the footnote using inline r code.

Table 11:

	Dependent variable:	
	age	
	(1)	(2)
shuckedWeight	6.111***	3.848***
	(0.204)	(0.230)
sexI		−2.249***
		(0.124)
sexM		−0.375***
		(0.106)
Constant	9.243***	10.914***
	(0.086)	(0.129)
Observations	4,169	4,169
R^2	0.177	0.243
Adjusted R^2	0.177	0.242
Residual Std. Error	2.924 (df = 4167)	2.805 (df = 4165)
F Statistic	897.663*** (df = 1; 4167)	445.716*** (df = 3; 4165)
Note:		*p<0.1; **p<0.05; ***p<0.01

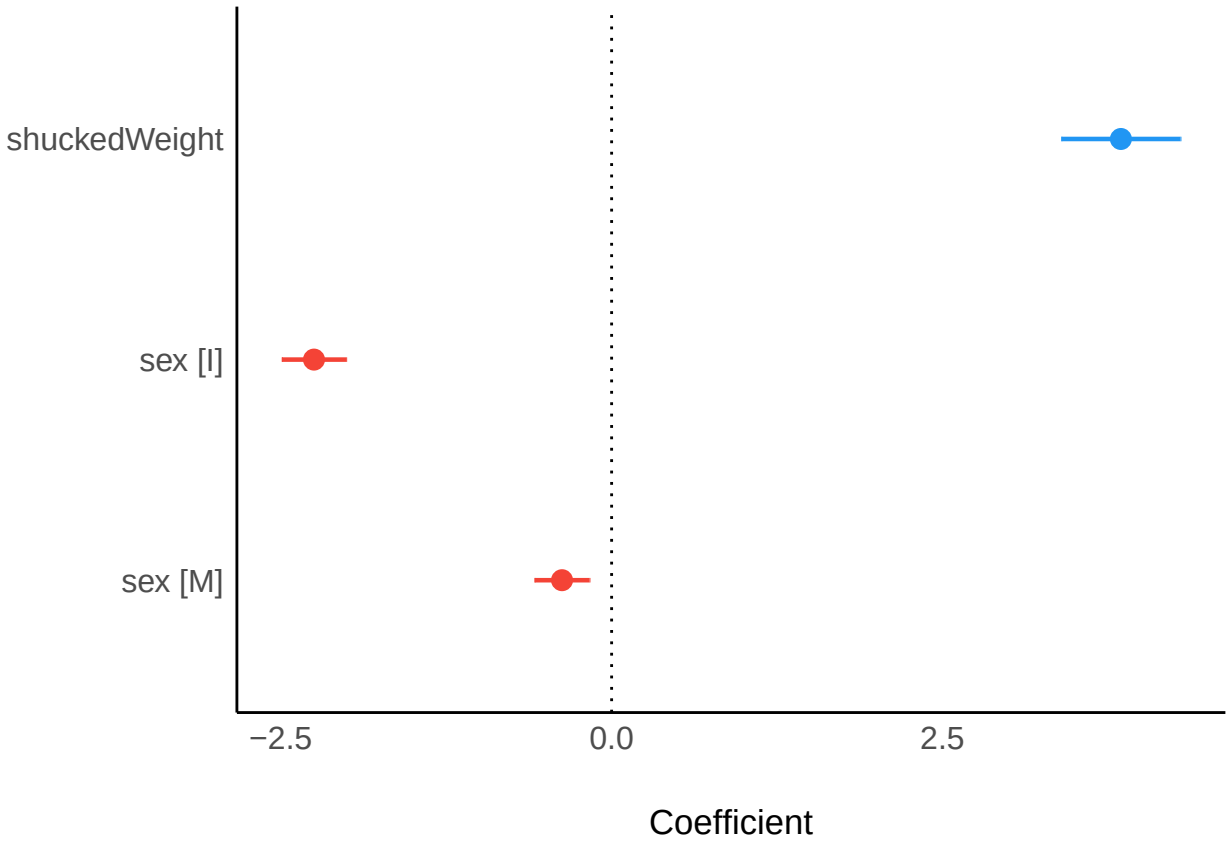
Predictor	b	b_95%_CI	sr2	sr2_95%_CI	Fit	Difference
(Intercept)	9.24**	[9.07, 9.41]				
shuckedWeight	6.11**	[5.71, 6.51]	.18	[.16, .20]		
					R2 = .177**	
					95% CI[.16,.20]	
(Intercept)	10.91**	[10.66, 11.17]				
shuckedWeight	3.85**	[3.40, 4.30]	.05	[.04, .06]		
sexI	-2.25**	[-2.49, -2.01]	.06	[.05, .07]		
sexM	-0.37**	[-0.58, -0.17]	.00	[-.00, .00]		
					R2 = .243**	Delta R2 = .066**
					95% CI[.22,.26]	95% CI[.05, .08]

Note. A significant b-weight indicates the beta-weight and semi-partial correlation are also significant. b represents unstandardized regression weights. beta indicates the standardized regression weights. sr2 represents the semi-partial correlation squared. r represents the zero-order correlation. Square brackets are used to enclose the lower and upper limits of a confidence interval. * indicates p < .05. ** indicates p < .01.

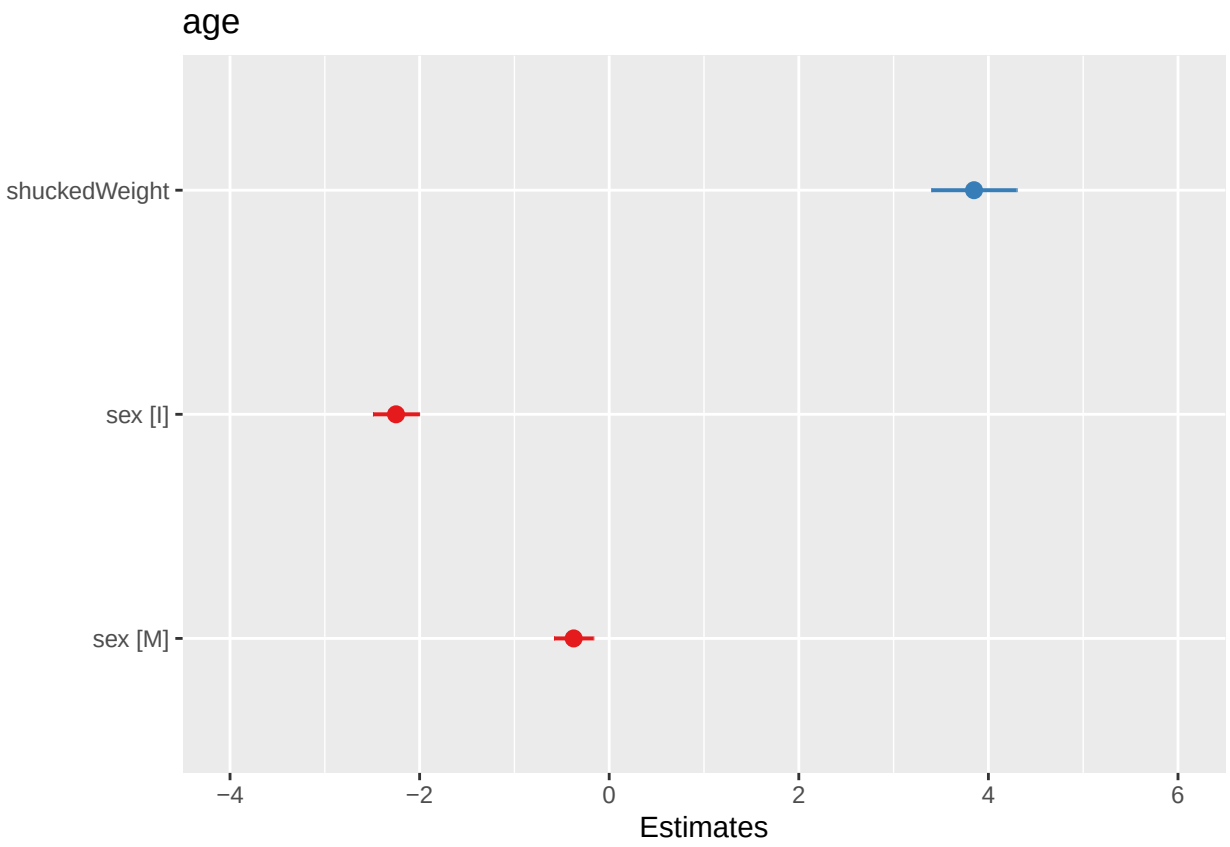
Visualize Regression Coefficients

Here is an example plot of the coefficients from Model 2 (`lm2`) using the packages from `easystats` - namely the `parameters` and `see` packages.

Learn more at https://easystats.github.io/easystats/.



Here is another example using the `sjPlot` package. Note: I also had to install/update the associated `sjstats` package. Learn more at http://www.strengejacke.de/sjPlot/reference/plot_model.html.



Logistic Regression of `adult` by Shucked Weight and Diameter

NOTE: The `adult` variable is currently a "character" class variable. So, let's create a 0/1 coded variable.

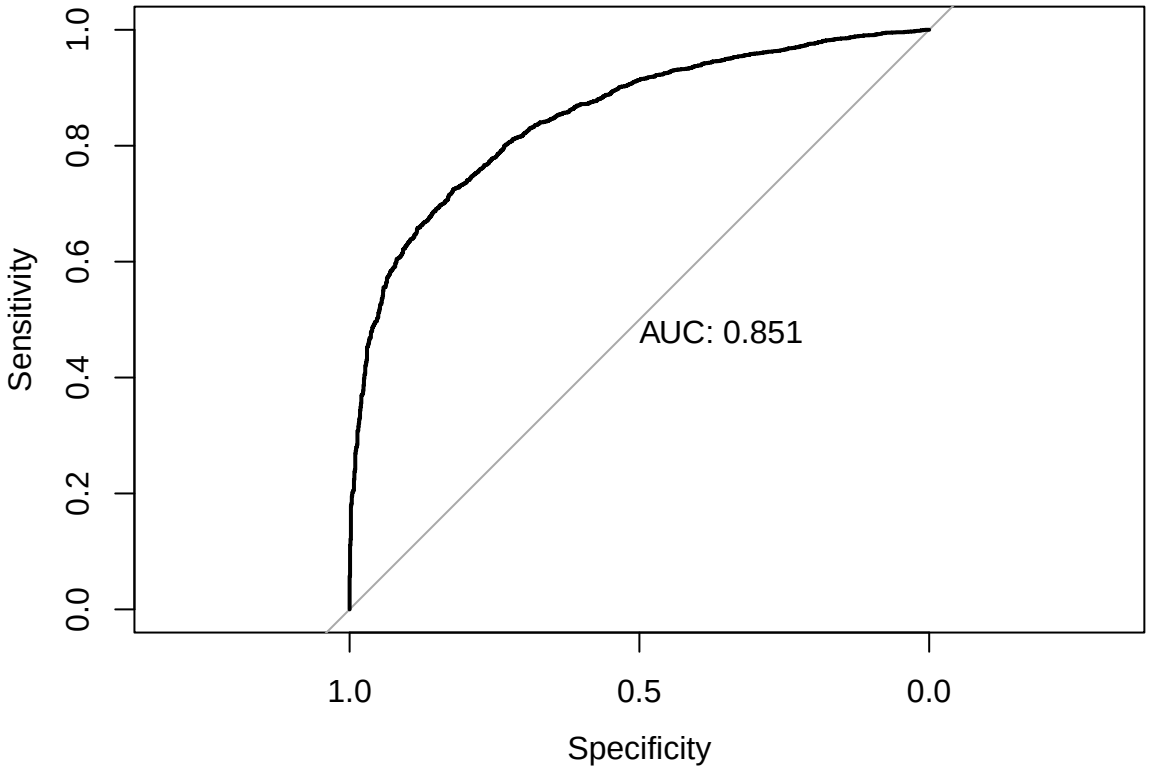
Use `parameters` package to get model coefficients table.

Parameter	Coefficient	SE	CI	CI_low	CI_high	z	df_error	p
(Intercept)	0.0365486	0.0110825	0.95	0.0200387	0.0657934	-10.913014	Inf	0e+00
shuckedWeight	217.0126805	139.5074649	0.95	62.5988462	778.3652734	8.368861	Inf	0e+00
diameter	540.2870048	637.2057215	0.95	53.9313991	5495.9308663	5.335075	Inf	1e-07

Another option using the `gtsummary` package.

Characteristic	**OR**	**95% CI**	**p-value**
shuckedWeight	217	62.6, 778	<0.001
diameter	540	53.9, 5,496	<0.001

Compute the AUC for the model and plot the ROC curve.




```
##
## Call:
## roc.formula(formula = abalone$adult01 ~ pred_glm1, plot = TRUE,     print.auc = TRUE)
##
## Data: pred_glm1 in 1335 controls (abalone$adult01 0) < 2834 cases (abalone$adult01 1).
## Area under the curve: 0.8507
```