L11 MLR – Abalone Quiz

ENVX1002 Statistics in Life and Environmental Sciences

Si Yang Han

The University of Sydney

Feb 2026

Abalone Quiz

This dataset records abalone from the coast of Tasmania, Australia (Nash, 1995) and was accessed from the [UCI

Introduction

Abalone are marine snails that are a considered a delicacy and very expensive. The older the abalone, the higher the price. Age is determined by counting the number of rings in the shell. To do this, the shell needs to be cut, stained and viewed under a microscope - which is a lot of effort. Researchers measured 9 attributes of the abalone: `sex`, `length`, `diameter`, `height`, `whole`, `shucked`, `viscera`, `shell`, and `rings`.

Note: `whole`, `shucked`, `viscera` and `shell` are weight measurements.

. . .

What is the response variable?

a) length
b) rings
c) shell (weight)
d) whole (weight)

. . .

> Reading comprehension :)

Research question

Abalone are marine snails that are a considered a delicacy and very expensive. The older the abalone, the higher the price. Age is determined by counting the number of rings in the shell. To do this, the shell needs to be cut, stained and viewed under a microscope - which is a lot of effort. Researchers measured 9 attributes of the abalone: `sex`, `length`, `diameter`, `height`, `whole`, `shucked`, `viscera`, `shell`, and `rings`.

Note: `whole`, `shucked`, `viscera` and `shell` are weight measurements.

What is the best research question, based on the context above?

a) Is there a correlation between abalone age and weight?
b) Can abalone weight be predicted from other measured variables?
c) Is there a relationship between abalone size and age?
d) Can age be measured by size?

…

> A is not a complete answer, there are many more predictors. B is incorrect, we care about age/rings. D is incorrect, we are trying to *model or predict abalone age from size* – terminology matters.

Explore data

We sample the data to make it easier to visualise relationships. We also remove the `sex` variable because it is not numeric.

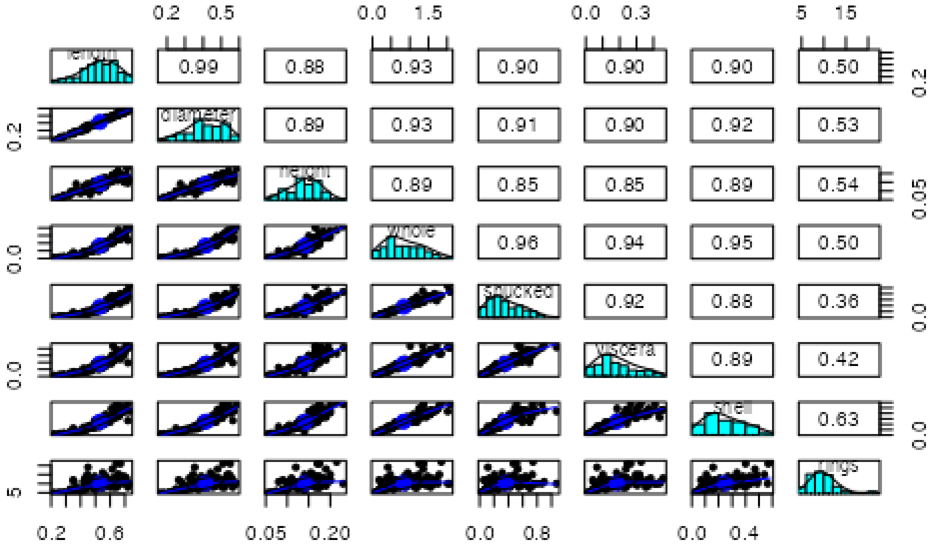
```r
abalone ← read.csv("data/abalone.csv")

set.seed(1113)          # reproducible randomness
abalone ← abalone %>%
  select(-sex) %>%      # remove `sex` because it is categorical
  sample_n(100)         # sample 100 observations for cleaner curve

str(abalone)
```

```
'data.frame':   100 obs. of  8 variables:
 $ length  : num  0.52 0.71 0.33 0.67 0.65 0.35 0.695 0.52 0.6 0.61 ...
 $ diameter: num  0.405 0.57 0.255 0.55 0.51 0.25 0.53 0.41 0.475 0.48 ...
 $ height  : num  0.14 0.195 0.095 0.17 0.19 0.1 0.15 0.14 0.15 0.17 ...
 $ whole   : num  0.692 1.348 0.188 1.247 1.542 ...
 $ shucked : num  0.276 0.8985 0.0735 0.472 0.7155 ...
 $ viscera : num  0.137 0.444 0.045 0.245 0.373 ...
 $ shell   : num  0.215 0.454 0.06 0.4 0.375 ...
 $ rings   : int  11 11 7 21 9 7 14 11 10 10 ...
```

```
psych::pairs.panels(abalone)     # visualise relationships
```



What is the most correlated predictor with (number of) abalone rings?

a) age
b) length
c) whole (weight)
d) shell (weight)

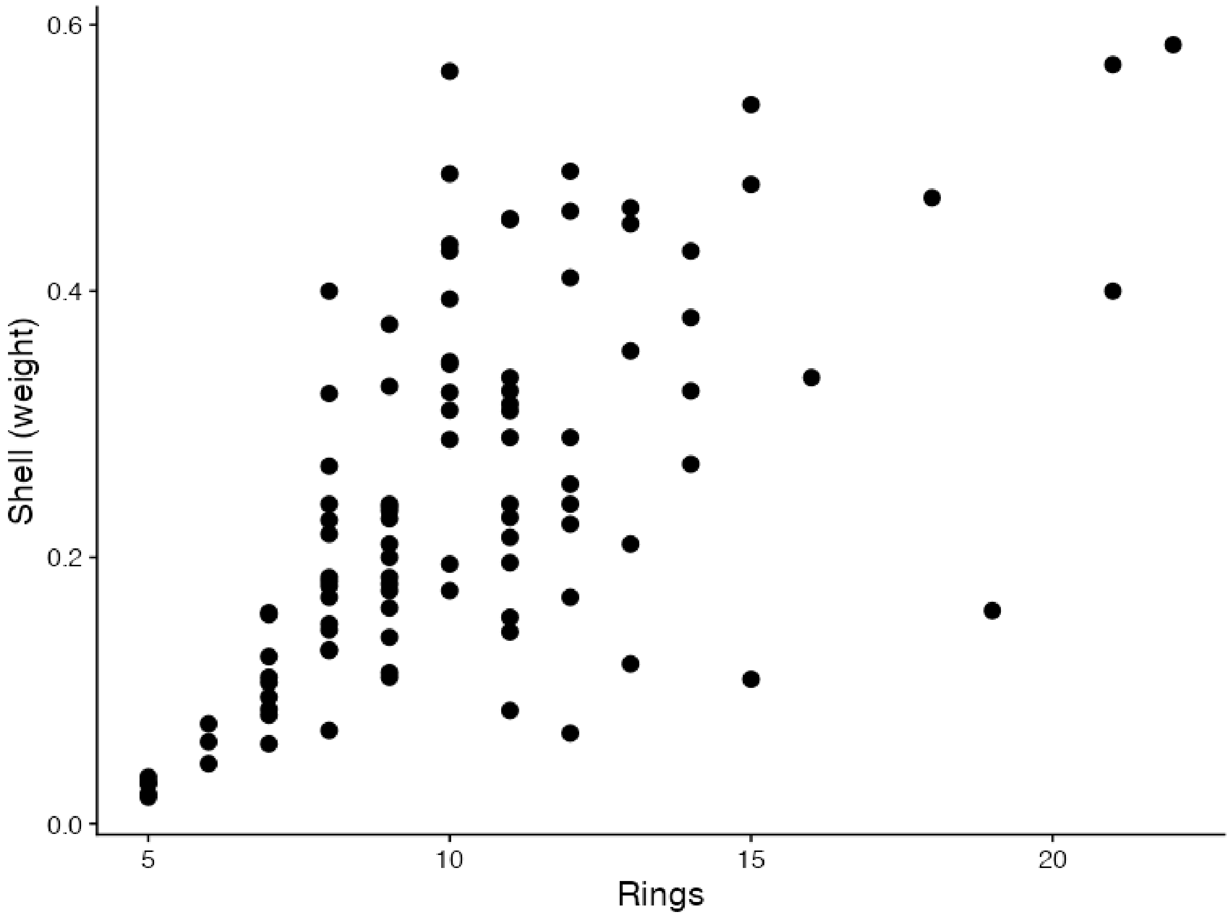
> Age is a trick - it is not a predictor! The answer is `shell` (weight).

```r
ggplot(abalone, aes(x = rings, y = shell)) +
  geom_point(size = 3) +
  labs(x = "Rings", y = "Shell (weight)") +
  theme(text = element_text(size = 16))
```

Which assumption/s do we need to be wary of?

a) linearity
b) collinearity
c) equal variances
d) all of the above

> In the `ring` plot, the relationship is not clearly linear and there is fanning (unlikely equal variances met). There is also very high correlation between some predictors (e.g. `length` and `diameter`), so the answer is D.


0.6
0.4
0.2
0.0
Shell (weight)
5
10
15
20
Rings

```
cor(abalone) |> round(2) |> print() # visualise
relationships
```

```
         length diameter height whole shucked viscera
shell rings
length     1.00     0.99   0.88  0.93    0.90    0.90
0.90  0.50
diameter   0.99     1.00   0.89  0.93    0.91    0.90
0.92  0.53
height     0.88     0.89   1.00  0.89    0.85    0.85
0.89  0.54
whole      0.93     0.93   0.89  1.00    0.96    0.94
0.95  0.50
shucked    0.90     0.91   0.85  0.96    1.00    0.92
0.88  0.36
viscera    0.90     0.90   0.85  0.94    0.92    1.00
0.89  0.42
shell      0.90     0.92   0.89  0.95    0.88    0.89
1.00  0.63
```

```
rings      0.50     0.53   0.54  0.50    0.36    0.42
0.63  1.00
```

Fit a model

We use natural log transformation on the response variable with `log()` to account for non-linear relationships.

```r
fit ← lm(log(rings) ~ ., data = abalone)
summary(fit)
```

```

Call:
lm(formula = log(rings) ~ ., data = abalone)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37297 -0.12727 -0.01584  0.08787  0.61636 

Coefficients:
            Estimate Std. Error t value Pr(>|
t|)    
(Intercept)  1.34626    0.18219   7.389 6.57e-11
***
```

Which predictor is NOT significant to the model?

a) height
b) whole (weight)
c) shucked (weight)
d) viscera (weight)

> Whole (weight) has a period (.) beside the p-value – this means the value is less than 0.10, but it needs to be <0.05 to be considered significant.

```
length      -1.25389    1.50969  -0.831
0.40837
diameter     3.24138    1.91481   1.693
0.09388 .
height       2.26408    1.34813   1.679
0.09646 .
whole        0.03089    0.29250   0.106
0.91612
shucked     -1.30902    0.38861  -3.368  0.00111
**
viscera     -0.24785    0.55098  -0.450
0.65389
shell        1.73328    0.60179   2.880  0.00494
**
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1

Residual standard error: 0.1996 on 92 degrees of
freedom
Multiple R-squared:  0.6187,    Adjusted R-
```

```
squared:  0.5897
F-statistic: 21.32 on 7 and 92 DF,  p-value: <
2.2e-16
```

Fit a model

```
Residual standard error: 0.1996 on 92 degrees of freedom
Multiple R-squared:  0.6187,    Adjusted R-squared:  0.5897
F-statistic: 21.32 on 7 and 92 DF,  p-value: < 2.2e-16
```

We determine model fit with:

a) Multiple R-squared and p-value
b) Adjusted R-squared and p-value
c) Adjusted R-squared and residual standard error
d) Multiple R-squared and residual standard error

. . .

> We have multiple variables, so we use the Adjusted R-squared. The p-value tests the hypothesis on whether the model should be used at all, in favour of the mean. The residual error is a measure of model fit.

The problem with using too many predictors

Here, the model is fit with all predictors, then the least significant predictor is removed. This process is repeated until only one predictor remains.

```r
library(broom)

full7 <- lm(log(rings) ~ ., data = abalone)
part6 <- update(full7, . ~ . - whole)
part5 <- update(part6, . ~ . - viscera)
part4 <- update(part5, . ~ . - length)
part3 <- update(part4, . ~ . - height)
part2 <- update(part3, . ~ . - diameter)
part1 <- update(part2, . ~ . - shucked)

formulas <- c(part1$call$formula,
              part2$call$formula,
              part3$call$formula,
              part4$call$formula,
              part5$call$formula,
              part6$call$formula,
               full7$call$formula)

rs <- bind_rows(glance(part1),
                glance(part2),
                glance(part3),
                glance(part4),
                glance(part5),
                glance(part6),
                glance(full7)) %>%
  mutate(Model = formulas, n = 1:7) %>%
  select(Model, n, r.squared, adj.r.squared) %>%
  mutate_if(is.numeric, round, 3)

knitr::kable(rs)
```

Model	n	r.squared	adj.r.squared
log(rings) ~ shell	1	0.445	0.439
log(rings) ~ shucked + shell	2	0.557	0.548
log(rings) ~ diameter + shucked + shell	3	0.604	0.591
log(rings) ~ diameter + height + shucked + shell	4	0.614	0.598

Model	n	r.squared	adj.r.squared
log(rings) ~ length + diameter + height + shucked + shell	5	0.618	0.597
log(rings) ~ length + diameter + height + shucked + viscera + , shell	6	0.619	0.594
log(rings) ~ .	7	0.619	0.590

Considering only R^2, which model would we choose?

a) Model with 1 predictor
b) Model with 3 predictors
c) Model with 4 predictors
d) Model with 7 predictors

> The 1-predictor model sacrifices 14.5% of variation in the response (too much). The 7-predictor model is overfitted (worse than 4-predictor model). Between 3 and 4-predictor models - is a 0.7% improvement worth having to measure `height`? Realistically, the models with 2 or 3 predictors are justifiable.

Interpretation

```
#| eval: false
Call:
lm(formula = log(rings) ~ diameter + shucked + shell, data = abalone)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30290 -0.15469 -0.03485  0.11454  0.64573

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4122     0.1594   8.859 4.19e-14 ***
diameter      2.0346     0.6034   3.372  0.00108 **
shucked      -1.3339     0.2152  -6.200 1.42e-08 ***
shell         2.0486     0.3672   5.579 2.23e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Which is the correct equation?‘

a) `log(rings)` = 1.41 + 2.04 x `diameter` - 1.33 x `shucked` + 2.04 x `shell`

> Attention to detail :)

b) `log(rings)` = 1.41 + 2.03 x `diameter` - 1.33 x `shucked` + 2.04 x `shell`
c) `log(rings)` = 1.41 + 2.04 x `diameter` + 1.33 x `shucked` + 2.05 x `shell`
d) `log(rings)` = 1.41 + 2.03 x `diameter` - 1.33 x `shucked` + 2.05 x `shell`

Interpretation

The equation of our model is:

`log(rings)` **= 1.41 + 2.03 x** `diameter` **+ −1.33 x** `shucked` **+ 2.05 x** `shell`

Below are three statements. *Given all other predictors are held constant:*

- `rings` changes by $e^{-1.33}$ for every percent increase in `shucked` (weight)
- `log(rings)` changes by 1.33 for every unit increase in `shucked` (weight)
- `log(rings)` changes by approximately 1.33% for every percent increase in `shucked` (weight)

How many statements are correct?

a) none
b) 1 statement
c) 2 statements
d) all of them

> The first two are correct, the third is not. The natural log percent change appoximation only applies to small β values below $|0.25|$.

The most important question

How do you feel about regression so far?

a) Easy
b) OK
c) Hard
d) SOS

Good work!